Universidade Federal de Sergipe

Pró-Reitoria de Pós-Graduação e Pesquisa

Programa de Pós-Graduação em Engenharia Elétrica

Estudo Dimensional de Características Aplicadas à Leitura Labial Automática

Dissertação de Mestrado

Fillipe Levi Guedes Madureira

Orientador: Jugurta Rosa Montalvão Filho

São Cristóvão, SE – Brasil

Agosto de 2018

Fillipe Levi Guedes Madureira

Estudo Dimensional de Características Aplicadas à Leitura Labial Automática

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia Elétrica – PROEE, da Universidade Federal de Sergipe, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia Elétrica.

Universidade Federal de Sergipe

Pró-Reitoria de Pós-Graduação e Pesquisa

Programa de Pós-Graduação em Engenharia Elétrica

Orientador: Jugurta Rosa Montalvão Filho

São Cristóvão, SE – Brasil

Agosto de 2018

UNIVERSIDADE FEDERAL DE SERGIPE
PRÓ-REITORIA DE PÓS-GRADUAÇÃO E PESQUISA
COORDENAÇÃO DE PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA-PROEE

TERMO DE APROVAÇÃO

"Estudo dimensional de características aplicadas à leitura labial"

Discente:

Fillipe Levi Guedes Madureira

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Sergipe, como parte dos requisitos exigidos para a obtenção do título de Mestre em Engenharia Elétrica.

Aprovada pela banca examinadora composta por:

Prof. Dr. Eduardo Oliveira freire (PROEE/UFS)
Presidente

Prof. Dr. Lucas Molina (DEL/UFS)
Examinador Externo

Prof. Dr. Jânio Coutinho Canuto (PROEE/UFS)
Examinador Interno

Fillipe Levi Guedes Madureira
Fillipe Levi Guedes Madureira
Candidato

Cidade Universitária "*Prof. José Aloísio de Campos*", 31 de agosto de 2018.

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL
UNIVERSIDADE FEDERAL DE SERGIPE

Madureira, Fillipe Levi Guedes
M183e Estudo dimensional de características aplicadas à leitura labial automática / Fillipe Levi Guedes Madureira ; orientador Jugurta Rosa Montalvão Filho . - São Cristóvão, 2018.
64 f. : il.

Dissertação (mestrado em Engenharia Elétrica) – Universidade Federal de Sergipe, 2018.

1. Engenharia elétrica. 2. Surdos - Meios de comunicação. 3. Comunicação oral 4. Fala. 5. Sistemas de reconhecimento de padrões I. Montalvão Filho, Jugurta Rosa orient. II. Título.

CDU 621.3:81'1

Agradecimentos

Agradeço à minha família pelo carinho e apoio incondicional em todo o meu trajeto existencial.

Agradeço especialmente ao meu orientador, Jugurta Montalvão, pelos ensinamentos valiosos, paciência, incentivos, compreensão e inspiração dedicados a mim. Agradeço, também, aos professores com os quais tive contato durante o curso de mestrado e me ajudaram direta ou indiretamente.

Agradeço aos meus amigos que se dispuseram a ouvir-me falar sobre este trabalho, participaram de experimentos, demonstraram entusiasmo e me incentivaram, em especial a Taís Calado, Renan Silva e Filipe Barreto.

Por fim, agradeço à CAPES (Coordenação de Aperfeiçoamento Pessoal de Nível Superior) pelo apoio financeiro, sem o qual seria impossível concluir este trabalho.

Resumo

Este trabalho é um estudo da relação entre a dimensão intrínseca de vetores de características aplicados à classificação de sinais de vídeo no intuito de realizar-se a leitura labial. Nas tarefas de reconhecimento de padrões, a extração de características relevantes é crucial para um bom desempenho dos classificadores. O ponto de partida deste trabalho foi a reprodução do trabalho de J.R. Movellan [1], que realiza a classificação de gestos labiais com HMM na base de dados Tulips1, utilizando somente o sinal de vídeo. A base é composta por vídeos das bocas de voluntários enquanto esses pronunciam os primeiros 4 numerais em inglês. O trabalho original utiliza vetores de características de dimensão muito alta em relação ao tamanho da base. Consequentemente, o ajuste de classificadores HMM se tornou problemático e só se alcançou 66,67% de acurácia. Estratégias de extração de características e esquemas de classificação alternativos foram propostos, a fim de analisar a influência da dimensão intrínseca no desempenho de classificadores. A melhor solução, em termos de resultados, obteve uma acurácia de aproximadamente 83%.

Palavras-chave: dimensão intrínseca, extração de características, leitura labial, HMM.

Abstract

This work is a study of the relationship between the intrinsic dimension of feature vectors applied to the classification of video signals in order to perform lip reading. In pattern recognition tasks, the extraction of relevant features is crucial for a good performance of the classifiers. The starting point of this work was the reproduction of the work of J.R. Movellan [1], which classifies lips gestures with HMM using only the video signal from the Tulips1 database. The database consists of videos of volunteers' mouths while they utter the first 4 numerals in English. The original work uses feature vectors of high dimensionality in relation to the size of the database. Consequently, the adjustment of HMM classifiers has become problematic and the maximum accuracy was only 66.67%. Alternative strategies for feature extraction and classification schemes were proposed in order to analyze the influence of the intrinsic dimension in the performance of classifiers. The best solution, in terms of results, achieved an accuracy of approximately 83%.

Keywords: intrinsic dimension, feature extraction, lip-reading, HMM.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

AAM	*Active Appearance Model*
ASM	*Active Shape Model*
ASR	*Automatic Speech Recognition*
AVASR	*Audiovisual Automatic Speech Recognition*
BRIEF	*Binary Robust Independent Elementary Features*
CNN	*Convolutional Neural Network*
DI	Dimensão intrínseca
DTW	*Dynamic Time Warping*
EM	*Expectation Maximization*
fps	*Frames* por segundo
HMM	*Hidden Markov Model*
PCA	*Principal Component Analysis*
SVM	*Support Vector Machine*

Lista de símbolos

σ_f	Desvio padrão do filtro gaussiano
$\mathbf{S}_f$	Vetor de coordenadas dos pontos de referência dos lábios
N_p	Número de pontos de referência rastreados
P_r	r-ésimo ponto de referência rastreado
$\mathbf{V}_{r,t}$	Vetor descritor do r-ésimo ponto rastreado no *frame* t
$\mathbf{g}_k$	Vetor de características do k-ésimo gesto por concatenação de 3 imagens em nível de cinza
$\mathbf{b}_k$	Vetor de características do k-ésimo gesto por concatenação de 3 imagens em nível de cinza codificadas com BRIEF
$\mathbf{p}_k$	Vetor de características do k-ésimo gesto por concatenação das coordenadas dos pontos de referência rastreados

Sumário

1 Introdução

1.1 A Natureza Bimodal da Fala

A fala humana tem natureza bimodal. Os seres humanos combinam, sempre que possível, informações auditivas e visuais ao interpretar e decidir sobre o que foi dito. Aqueles com deficiências auditivas são capazes de compreender a fala fluentemente através de leitura labial, ou seja, através do processamento da informação visual obtida dos lábios e face do orador. Mesmo para os não deficientes auditivos, é sabido que a simples visualização do rosto do orador aumenta significativamente a inteligibilidade, especialmente sob condições ruidosas [2, 3].

A inteligibilidade é beneficiada pelas informações visuais, que proveem informações acerca da localização do orador (fonte sonora) e da segmentação de fala. Os ganhos de inteligibilidade são atribuídos à presença de informação complementar ao áudio, referente ao lugar de articulação da fala produzida, ou seja, são devidos à visibilidade parcial ou total dos articuladores e ao limite de terminação do trato vocal, em relação à colocação da língua, visibilidade dos dentes, formato dos lábios e boca e, em menor grau, movimento do queixo e bochechas [4]. Tais informações podem ajudar a dirimir ambiguidades no domínio sonoro. Por exemplo, as consoantes 'b' e 'v', embora similares no domínio acústico, têm representações visuais distintas.

A natureza bimodal da fala pode ser demonstrada por meio do Efeito McGurk [5, 6], que ocorre quando o indivíduo 'ouve' algo diferente do que foi dito devido à influência de um estímulo visual conflitante. Nos experimentos de McGurk e MacDonald[5], quando o som '*ba*' foi superposto ao vídeo de uma mulher pronunciando repetidamente a sílaba '*ga*', a maioria das pessoas testadas, adultos saudáveis, percebeu uma terceira sílaba '*da*'. Esse fenômeno ilustra a complexidade do mecanismo de fusão de informação bimodal que existe no cérebro para percepção da fala.

As representações auditiva e visual têm uma certa sincronia e são correlacionadas entre si, uma vez que ambos os sinais são gerados conjuntamente na cavidade orofacial. Isso permite que um seja utilizado para recuperar, parcialmente, a representação do outro, característica útil para inúmeras aplicações bimodais como, por exemplo, sintetizar voz artificialmente a partir de vídeos [7]. Quanto à sincronicidade, frequentemente o sinal de áudio pode se manisfestar atrasado em relação ao visual, devido ao prévio posicionamento dos articuladores em preparação para a pronunciação de um determinado som [4].

Diversos estudos sobre a percepção da fala e leitura labial foram conduzidos no século XX e, embora já houvesse hipóteses sobre a relação díspar entre sons (fonemas) e

suas correspondentes manifestações visuais através dos lábios, o estudo de Fisher [8] ficou conhecido por cunhar o termo *visema*. Os fonemas são as representações das unidades sonoras, as partículas que constituem e diferenciam palavras. Os visemas são seus análogos, as unidades básicas distinguíveis entre si no domínio da percepção visual. Consistem, basicamente, de agrupamentos das unidades básicas de som, ou seja, são derivados de fonemas que manifestam a mesma aparência visual (abordagem linguística). Contudo, os fonemas são produzidos pela combinação de diversos fatores, i.e., posicionamento e movimento dos articuladores como as cordas vocais, língua, palato mole etc, dos quais apenas alguns poucos são visíveis. Portanto, é natural que a relação entre fonemas e visemas seja mapeada de vários para um, havendo muitos fonemas visualmente indistinguíveis. Aliás, não há sequer um consenso na literatura quanto ao mapeamento entre fonemas e visemas [4].

Há uma outra maneira de definir os visemas, através de uma abordagem orientada pelos dados, que é baseada nos gestos articulatórios, tais como o fechamento ou arredondamento dos lábios, exposição dos dentes, movimento da mandíbula etc, sem estabelecer, porém, uma conexão com o fonema produzido. Nenhuma das definições sugere, contudo, que os visemas sejam capazes de distinguir palavras da mesma forma que os fonemas [9]. De fato, a quantidade de informações de fala disponível no sinal visual é significativamente menor do que num sinal de áudio sem ruído [4].

1.2 O Reconhecimento de Fala Como Ciência

O reconhecimento automático de fala ou ASR – do inglês *Automatic Speech Recognition* – é uma área de pesquisa que tem sido desenvolvida desde o início da década de 1950. Na época, surgiram os primeiros trabalhos para reconhecimento de dígitos isolados [10], 10 sílabas de um único orador [11] e 10 vogais [12]. Nas décadas seguintes, viu-se um aumento progressivo do tamanho dos vocabulários e diminuição nas taxas de erro no reconhecimento de palavras devido ao desenvolvimento e aplicação de novas técnicas e algoritmos. Os grandes avanços aconteceram nas décadas de 1960 e 1970, com a introdução de técnicas de representação da fala avançadas baseadas em LPC (*Linear Predictive Coding*) e análise cepstral, e na década de 1980 através da introdução de modelos estatísticos rigorosos baseados em HMM (*Hidden Markov Model*) [13]. Na década de 1990, as tecnologias de ASR finalmente atingiram o mercado. Surgiram as primeiras aplicações comerciais, empregadas em serviços de atendimento ao consumidor e *call centers*.

Embora o reconhecimento de fala fosse uma matéria amplamente estudada e já se soubesse da natureza bimodal da fala, o primeiro sistema a fazer uso de leitura labial só foi completado em 1984, 30 anos depois do primeiro reconhecedor acústico. Mais detalhes sobre o desenvolvimento e o estado atual das pesquisas em leitura labial automática serão

dados no Capítulo 2. Apesar do atraso, recentemente, um sistema de inteligência artificial superou o desempenho humano nessa tarefa [14], porém os sistemas que empregam leitura labial ainda não estão disponíveis para o grande público.

Atualmente, a ASR está presente em diversos sistemas móveis, de automação, assistentes virtuais, transcritores de texto etc. O campo de reconhecimento automático de fala tem se beneficiado dos avanços nas áreas de *deep learning* [15] e *big data*. Grandes empresas como o Google, Apple, IBM, Microsoft e outras têm utilizado essa enorme quantidade de dados disponível e novas técnicas de *deep learning* para criar sistemas mais robustos e empregá-los nos seus produtos.

Apesar de haver técnicas estabelecidas e consolidadas para a ASR e dos recentes avanços proporcionados pela grande disponibilidade de dados, o reconhecimento de fala ainda está longe de ser um campo de pesquisa estagnado e infrutífero. A maioria dos sistemas de ASR são desenvolvidos objetivando aplicações específicas e consideram um ambiente relativamente controlado. Os sistemas necessitam de melhorias em aspectos como a utilização por múltiplos usuários, sotaques, microfones, interferência de canal ou ambiente ruidoso [2]. Mesmo em situações completamente favoráveis, o estado da arte neste campo tecnológico não obtém um desempenho comparável à percepção humana.

O ímpeto pelo desenvolvimento de melhores sistemas de ASR combinado à natureza bimodal da fala motivam a pesquisa e o desenvolvimento de sistemas de ASR que façam uso de leitura labial automática – também chamados de sistemas de reconhecimento de fala multimodais, audiovisuais ou AVASR (*Audio-visual Automatic Speech Recognition*) – pois a integração de informações visuais nos sistemas baseados puramente em sinais acústicos aumenta o desempenho dos mesmos [16], especialmente sob condições ruidosas. Além disso, os AVASR permitem o reconhecimento quando o sinal de áudio estiver temporariamente inacessível ou mesmo na sua total ausência (sistemas de reconhecimento puramente visual de fala).

As aplicações dos sistemas de leitura labial são extremamente variadas. Podem permitir que pessoas com determinadas patologias vocais (indivíduos que sofreram laringectomia, por exemplo) possam se comunicar com maior naturalidade assim como permitir a comunicação confidencial ou de maneira a não incomodar outros ocupantes em locais públicos (por exemplo, em salas de reunião, bibliotecas, teatros etc). Outras aplicações são a televigilância e a extração de discurso para propósitos forenses [17, 18], além de facilitar a comunicação em ambientes ruidosos como em indústrias e *cockpits* de aeronaves. Chung et al.[14] salientam ainda a possibilidade de que a pesquisa na área de leitura labial possa discernir importantes sinais discriminativos que sejam benéficos no ensino de leitura labial às pessoas surdas ou com problemas de audição.

A leitura labial automática exige que características visuais que sejam informativas quanto ao discurso sejam extraídas do vídeo da face do orador. Um grande problema

na identificação das características é a enorme quantidade de dados contida em vídeos, um problema comum aos sistemas de visão computacional [2]. Zhou et al.[19] aborda os principais aspectos da extração de características para leitura labial.

> Idealmente, as características visuais deveriam ser relativamente compactas e suficientemente informativas em relação ao discurso proferido, ao mesmo tempo em que demonstram um certo nível de invariância às informações irrelevantes ou ruído nos vídeos. É um problema desafiador em grande parte devido aos fatos de haver incertezas (e.g., identidade do orador e pose da cabeça) que podem significativamente afetar a aparência visual da boca que fala em imagens e que as características visuais são extraídas para descrever um processo dinâmico (a pronunciação) em vez de imagens estáticas.

No domínio acústico, o MFCC (*Mel-frequency cepstral coefficients*) é uma ferramenta poderosa na representação de sons, uma vez que imita parcialmente a percepção humana dos sons. Desde sua concepção, permanece praticamente inalterado [20]. Em contraste, na área de leitura labial automática ainda não há uma técnica para extração de características universalmente aceita para representar o sinal visual da fala apesar das mais de 3 décadas de pesquisa.

1.3 Objetivos

O primeiro objetivo desta dissertação é realizar o reconhecimento de fala independentemente da mudança do orador a partir de pontos de referência rastreados na região dos lábios, considerando um vocabulário limitado composto pelos numerais e sem continuidade, ou seja, a expressão oral de cada palavra é feita de forma isolada, fora do contexto de uma oração. Para tal, serão desconsiderados quaisquer sinais acústicos e utilizados apenas os de vídeo.

O segundo objetivo deste trabalho é realizar a estimação da dimensão intrínseca dos vetores de características extraídos e analisar seu impacto nas acurácias dos classificadores, considerando diversos esquemas de classificação.

1.4 Organização do Trabalho

Esta dissertação está dividida da seguinte forma: no Capítulo 2 encontra-se uma revisão sobre o problema de reconhecimento de fala; no Capítulo 3, faz-se uma explanação da arquitetura de sistemas de leitura labial e dos outros conceitos teóricos abordados neste trabalho como dimensão intrínseca, codificação BRIEF e HMM; no Capítulo 4 são descritos todos os experimentos, além de análises sobre os resultados obtidos e; no Capítulo 5 são expostas as conclusões do trabalho.

2 Revisão Bibliográfica

O primeiro sistema de leitura labial foi completado em 1984 por Eric Petajan para sua tese de doutorado em engenharia elétrica, intitulada "*Automatic Lipreading to Enhance Speech Recognition*". O sistema utilizava quatro parâmetros da imagem da boca (altura, largura, perímetro e área) para realizar *template matching*. Os parâmetros eram derivados a partir de imagens binárias da boca, que eram registradas automaticamente através do rastreamento das narinas [21, 16, 22]. A evolução deste trabalho [23], empregava quantização vetorial das imagens antes do *template matching* com DTW (*Dynamic Time Warping*). Os experimentos utilizaram a pronunciação das letras do alfabeto e dos dígitos numéricos e a integração audiovisual foi realizada após o reconhecimento isolado de ambas as modalidades, através de um conjunto de regras heurísticas definidas *a priori*. Petajan mostrou que a combinação audiovisual obtinha desempenho superior ao de ambos os subsistemas isolados. Esses trabalhos mostraram o potencial da ASR audiovisual e motivaram a comunidade científica a realizar pesquisas nesse campo.

Em 1992, Hasegawa e Ohtani [24] realizaram uma das primeiras tentativas de sintetizar voz a partir de imagens da boca. A partir de parâmetros de baixo nível da imagem da boca (área, largura, altura) e língua (luminância média da área dos lábios), os autores estimaram a função de transferência do trato vocal e sintetizaram a voz com um filtro sintetizador PARCOR. O experimento consistiu em executar o áudio sintetizado e verificar se 17 indivíduos conseguiriam identificar as palavras corretamente. O vocabulário consistia das 5 vogais japonesas. A taxa de acerto média reportada foi de 91%.

Yuhas et al. [25] propuseram o uso de redes neurais para realizar a integração das modalidades de maneira mais "suave", se comparada ao método do segundo trabalho de Petajan et al. [23], já que as redes neurais são capazes de fundir diversas fontes de informação de maneira compacta e o algoritmo de treinamento elimina a necessidade do conjunto de regras a priori. O vocabulário consistiu de 9 sons vocálicos. Duas abordagens foram propostas: na primeira, que não utilizava fusão de modalidades, a rede neural mapeava diretamente as imagens estáticas de entrada nas vogais correspondentes; na segunda, estimativas independentes da função de transferência do trato vocal foram produzidas a partir dos sinais de áudio e das imagens da boca antes de serem combinadas por uma média ponderada para, então, serem submetidas ao sistema de reconhecimento. Uma rede neural foi treinada para mapear a imagem de entrada na estimativa do seu envelope do espectro acústico (*Short-Term Spectral Amplitude Envelope* – STSAE). A primeira abordagem atingiu uma taxa de acerto de 79% no conjunto de teste. A segunda obteve desempenho superior a 90%.

Wu et al. [26] também propuseram um sistema de ASR audiovisual baseado em redes neurais. Ambos os sinais foram inseridos simultaneamente, após devido pré-processamento, na entrada da rede. O vocabulário consistiu das vogais japonesas. O desempenho obtido com as modalidades combinadas foi 12% superior ao do áudio isolado. A taxa de reconhecimento utilizando apenas as características visuais foi de 50%.

Stork, Wolff e Levine [27] utilizaram uma abordagem diferente de Petajan et al. [23], que utilizou mapas de *pixels* de sequências inteiras de vídeo. Ao invés disso, dez marcadores foram fixados na face dos oradores e suas posições foram gravadas, resultando numa enorme redução da quantidade de informação processada. Essas alterações tornaram o sistema mais rápido e independente do orador. O sistema difere do proposto por Yuhas et al.[25], uma vez que não utiliza imagens estáticas, incorporando, portanto, variação temporal aos dados visuais. O classificador utilizado foi uma rede neural TDNN modificada (*Time Delay Neural Network*).

Bregler et al. [28] também utilizaram uma TDNN, porém incluíram DTW após as camadas das modalidades acústica e visual. O vocabulário consistiu de 26 fonemas do alfabeto alemão na forma de palavras e sequências aleatórias pronunciados continuamente. Os autores reportaram melhores resultados do que os obtidos por Stork, Wolff e Levine [27].

Goldschen, Garcia e Petajan [29] implementaram um sistema que tentava realizar o reconhecimento contínuo de fala utilizando apenas informações visuais. O objetivo era identificar corretamente as sentenças proferidas. O sistema codificava (por meio de um algoritmo de agrupamento) as características extraídas da imagem em vetores quantizados de um dicionário. As palavras do dicionário eram, então, apresentadas ao HMM para classificação. Este foi um dos primeiros trabalhos a utilizar HMM, numa tentativa de incorporar informações temporais.

Bregler e Konig [30] também usaram HMM como classificador. Além disso, fizeram uso de um tipo de modelo de contorno ativo (*active contour models*, popularmente conhecido como *snakes*) e aplicaram PCA (*Principal Component Analysis*) para reduzir a dimensionalidade das características. Eles denominaram essa abordagem de "*eigenlips*".

As *snakes* são *splines* (curvas definidas matematicamente por dois ou mais pontos de controle) que minimizam uma função de energia guiadas por forças externas restritivas e influenciadas por forças da imagem que puxam a curva na direção de características como linhas e bordas. Elas "travam" pontos de bordas próximas, localizando-as com precisão. São utilizadas em diversos problemas de visão computacional como detecção de bordas, linhas e contornos subjetivos e rastreamento de movimento [31]. Na formulação original, a energia tem um termo interno que deforma a curva na direção de pontos de interesse na imagem. São particularmente úteis para desenhar a forma de objetos amorfos. Entretanto, como nenhum modelo, exceto a suavização da curva, é imposto, as *snakes* não são ótimas

para localizar objetos com um formato definido [32]. Alguns trabalhos que utilizam *snakes* são os de Chiou e Hwang [33] e Chiou e Hwang [34].

Inspirados no trabalho de Kass, Witkin e Terzopoulos [31], Cootes e Taylor [35] desenvolveram os PDM (*Point Distribution Models*). Esses modelos são flexíveis, pois representam os objetos de interesse como conjuntos de pontos e fazem uma análise estatística das suas coordenadas em um diversos exemplos de treinamento. Esses modelos ficaram conhecidos como modelos de forma ativa (*Active Shape Models* – ASM) e se tornaram extremamente populares para extrair pontos de interesse na região da face, inspirando diversos trabalhos [36, 37, 38, 2].

Cootes, Edwards e Taylor [39, 40] descrevem outro modelo estatístico para modelar pontos de interesse, os modelos de aparência ativa (*Active Appearance Models* – AAM), como uma alternativa aos ASM. Segundo Cootes, Taylor et al. [32], os ASM descrevem a localização das estruturas numa imagem alvo, enquanto que os AAM manipulam um modelo capaz de sintetizar novas imagens do objeto de interesse, as quais podem ser usadas para classificação. Em [2, 41], os autores fazem uma comparação entre ASM e AAM.

Alguns trabalhos utilizaram preditores lineares como forma alternativa aos AAM para estimar a posição de pontos relevantes no contorno dos lábios [42, 43, 17]. Lan et al.; Lan, Harvey e Theobald [44, 45] usaram preditores lineares de forma complementar ao AAM, apenas para rastrear os pontos de referência a serem treinados no modelo de aparência. Bowden et al.; Bowden et al. [46, 18] usaram uma combinação de preditores lineares e AAM para treinar sistemas robustos quanto ao ângulo da câmera.

Graves et al. [47] apresentaram um método para rotular sequências de dados com redes neurais recorrentes (*Recurrent Neural Networks* – RNN) que eliminam a necessidade de dados de treinamento de entrada pré-segmentados e o pós-processamento das saídas (para transformá-las em sequências de rótulos), modelando todos os aspectos da sequência em uma mesma arquitetura. A rede interpreta as saídas como uma distribuição de probabilidades sobre todas as sequências de rótulos possíveis, dada uma determinada sequência de entrada. O método foi denominado de CTC (*Connectionist Temporal Classification*). Os resultados obtidos foram vantajosos em relação às implementações com apenas HMM ou com a abordagem híbrida HMM-RNN.

Noda et al. [48] propuseram a extração de características visuais para reconhecimento de fala utilizando uma rede neural convolucional (*Convolutional Neural Network* – CNN). Ao treinar a rede com imagens da região da boca do orador combinadas aos rótulos dos fonemas, a rede CNN obtém os filtros convolucionais essenciais para o reconhecimento dos fonemas. As dependências temporais das sequências de fonemas foram modeladas por um modelo HMM contínuo. O sistema foi avaliado em uma base de dados audiovisual formada por 300 palavras japonesas proferidas por 6 oradores diferentes. Os autores demonstram que as taxas de acerto desse sistema são superiores ao de sistemas que utilizam

abordagens de redução de dimensionalidade, incluindo PCA.

Wand, Koutník e Schmidhuber [3] propuseram um sistema de leitura labial baseado em redes neurais LSTM (*Long Short-Term Memory*) de ponta a ponta (*end-to-end*). O sistema foi avaliado e comparado experimentalmente a um classificador SVM (*Support Vector Machine*) que utilizou características convencionais no campo de visão computacional (*eigenlips* e histogramas). Os resultados obtidos (na base de dados GRID [49]) pela rede LSTM superaram aqueles do SVM, atingindo aproximadamente 80% em um vocabulário com 51 palavras.

Chung et al. [14] implementaram um sistema cujo objetivo é reconhecer frases e sentenças com ou sem o áudio disponível. O problema de reconhecimento de fala foi considerado aberto, ou seja, foram utilizadas sentenças de linguagem natural sem restrições, retiradas de vídeos da programação da televisão britânica. O sistema utiliza uma arquitetura de redes neurais denominada WLAS (*Watch, Listen, Attend and Spell*), que aprende a transcrever os vídeos dos movimentos das bocas em caracteres. O modelo WLAS treinado na base de dados criada pelos autores superou significativamente o desempenho de todos os trabalhos anteriores em outras bases de dados de referência.

3 Fundamentação Teórica

3.1 Os Sistemas de Leitura Labial Automática

Os sistemas de leitura labial são constituídos basicamente de três etapas: detecção e localização da região de interesse, extração de características e classificação/comparação dos gestos labiais.

3.1.1 Localização da Região de Interesse

Primeiramente, deve-se localizar a face do orador. Os primeiros esquemas de leitura labial, entretanto, utilizavam uma câmera fixa na face do orador, que permanecia relativamente estático. Algoritmos de detecção de face como o de Viola-Jones [50] e baseados em histogramas de gradientes (HOG) [51] são bastante populares. Alternativamente, redes CNN podem ser treinadas com esse intuito [15, 52].

Uma vez localizada a face, métodos de detecção de pontos de interesse na face (do inglês *facial landmark detection*) podem ser aplicados ao *bounding box* do detector de face. Esses pontos localizam estruturas como os olhos, canto dos lábios, narinas, entre outros, que ajudam a determinar a região da boca e a pose do rosto [53, 54, 55, 56].

Uma abordagem para a segmentação especificamente dos lábios é o uso de transformadas de cor. A segmentação é possível devido à diferenças de cores entre os lábios e pele no entorno [57]. Diversas transformadas aplicadas a este propósito foram comparadas em [58].

Para evitar a repetição desnecessária dos procedimentos de detecção de face e pontos de interesse, evitando o desperdício de recursos computacionais, algoritmos de rastreamento também são amplamente usados, sob a premissa de que o rosto e a localização de pontos de interesse não devem se descolocar significativamente entre dois *frames* consecutivos de um vídeo [4].

3.1.2 Extração de Características

Tradicionalmente há duas abordagens para a extração de características em um sistema de leitura labial: métodos baseados nos *pixels* da imagem e métodos baseados em modelos. Na primeira categoria, também chamada de abordagem *bottom-up* (de baixo para cima) as características são estimadas diretamente da imagem e exploram toda a informação em níveis de cinza da imagem. A segunda abordagem, conhecida como *top-down* (de cima para baixo) usa informações *a priori* e assunções encapsuladas num modelo que

descreve o contorno dos lábios. As características consistem, então, nos parâmetros do modelo ajustado à imagem [2, 57].

Métodos *Bottom-up*

1) Métodos baseados diretamente nos *pixels*: Wolff et al. [59] utilizou pontos manualmente marcados nas imagens como características; Movellan [1] usou imagens preprocessadas e Matthews et al. [2] extraiu características a partir de *sieves* (estrutura de filtro baseada em morfologia matemática).

2) Métodos baseados em transformações de imagens: Esta categoria faz uso de alguma transformação na região de interesse. Algumas transformadas utilizadas são a do cosseno discreta (DCT) [60], transformada Wavelet e de Fourier [61]. O PCA também é considerado nesta categoria por realizar redução dimensional baseada na variância dos dados.

3) Fluxo óptico: Este método é baseado na detecção de movimentos e assume que a informação sobre o movimento contém características relevantes [57]. Mase e Pentland [62] afirmam que o movimento muscular responsável pela articulação da fala também causa alterações na expressão facial e, portanto, deveria ser possível identificar a palavra proferida a partir do movimento muscular.

4) Métodos baseados em aprendizado profundo: Os trabalhos mais recentes na área têm utilizado redes neurais para a extração de características [48, 3, 63, 64]. A vantagem é que as redes conseguem extrair características invariantes à iluminação, translação e rotação, o que pode não acontecer com as técnicas anteriores.

Métodos *Top-down*

Esta abordagem descreve o contorno dos lábios com um número de parâmetros obtidos através da medição de posições, distâncias etc. A combinação linear dos parâmetros forma o vetor de características [57]. Os modelos desse tipo mais conhecidos são as *snakes*, propostas por Kass, Witkin e Terzopoulos [31] e os modelos ASM.

Métodos Híbridos

Combinam algumas vantagens de ambos os tipos de métodos descritos anteriormente para extrair características de movimento labial. Os modelos AAM [40] são a representação mais conhecida desta abordagem.

3.1.3 Classificadores e Comparadores

1) *Template matching*: Consiste na comparação de um vetor de características com uma referência. É um método que não considera a variação temporal das características, efetivamente comparando imagens estáticas, de implementação muito simples, mas que possui taxas de reconhecimento tipicamente baixas.

2) DTW: É um algoritmo utilizado para medir a similaridade e comparar sequências de vetores de dimensões (temporais) diferentes. É bastante utilizado em contextos de palavras isoladas, mas não obtém desempenho satisfatório em vocabulários extensos.

3) HMM: Modelo consagrado no campo da ASR tradicional, foi o sistema de classificação mais utilizado nas pesquisas até o recente crescimento da popularidade das redes neurais profundas, devido à natureza também estocástica dos movimentos labiais [57]. Mais detalhes sobre este modelo serão comentados na seção seguinte.

4) Redes neurais: Desde o início das pesquisas em AVASR, as redes neurais foram investigadas como ferramenta de classificação paralelamente ao HMM. Entretanto, sofreram com problemas de estimação de parâmetros. Além disso, não são estruturas adequadas para modelar dependências temporais. Contudo, com o desenvolvimento dos algoritmos de treinamento, maior disponibilidade de dados e computação paralela, as redes superaram essas barreiras. Frequentemente são utilizadas em arquiteturas híbridas, em conjunto com HMM, no qual as redes são responsáveis pelo papel das misturas gaussianas (GMM) [4].

3.2 Modelo Oculto de Markov

O modelo oculto de Markov – ou HMM (*Hidden Markov Model*) – é um modelo estatístico generativo utilizado para descrever sistemas como processos estocásticos. É amplamente utilizado para modelar sinais não estacionários, ou seja, sinais que apresentem evolução, dependência temporal.

No HMM, as observações são funções probabilísticas associadas aos estados. Estes são definidos por um processo estocástico, subjacente, que permanece oculto e somente pode ser observado através de outro processo estocástico que produz a sequência de observações [65]. Logo, é diferente da cadeia de Markov, na qual os estados são diretamente observados.

Existe uma probabilidade associada à mudança dos estados (ou permanência), que ocorre em intervalos regulares de tempo. Assim, os estados acumulam a informação da dinâmica temporal do sistema.

Considerando que um conjunto de estados é denotado por $S = \{S_1, S_2, S_3, \ldots, S_N\}$ e o estado no tempo t por q_t, um HMM com N estados escondidos é caracterizado por sua matriz de transição de estados $A = \{a_{ij}\}$, sendo

$$a_{ij} = P(q_t = S_i | q_{t-1} = S_j), \qquad 1 \leq i, j \leq N, \tag{3.1}$$

cujos coeficientes de transição obedecem as propriedades

$$a_{ij} \geq 0 \tag{3.2a}$$

e

$$\sum_{j=1}^{N} a_{ij} = 1, \tag{3.2b}$$

pela matriz de distribuição de probabilidade inicial dos estados $\pi = \{\pi_i\}$, sendo

$$\pi_i = P(q_1 = S_i), \qquad 1 \leq i \leq N, \tag{3.3}$$

e pelas distribuições de probabilidade de observação por estado $B = \{b_j(\mathbf{O})\}$, $1 \leq j \leq N$. As observações geradas pelo HMM podem ser discretas ou contínuas, a depender das definições de $b_j(\mathbf{O})$.

As observações contínuas são geralmente modeladas por uma mistura de variáveis aleatórias com distribuição de probabilidade normal, conforme

$$b_j(\mathbf{O}) = \sum_{m=1}^{M} c_{jm} \mathcal{N}(\mathbf{O}, \mu_{jm}, \mathbf{\Sigma}_{jm}) \tag{3.4}$$

onde $\mathbf{O}$ é o vetor de observação sendo modelado, μ_{jm}, $\mathbf{\Sigma}_{jm}$ e c_{jm} são, respectivamente, o vetor de médias, a matriz de covariância e o coeficiente da m-ésima componente da mistura no estado j. Os coeficientes da mistura devem obedecer às restrições

$$\sum_{m=1}^{M} c_{jm} = 1, \qquad 1 \leq j \leq N \tag{3.5a}$$

$$c_{jm} \geq 0, \qquad 1 \leq j \leq N, \quad 1 \leq m \leq M \tag{3.5b}$$

de forma que a distribuição de probabilidade seja propriamente normalizada, i.e.,

$$\int_{-\infty}^{\infty} b_j(\mathbf{O})\, d\mathbf{O} = 1, \quad 1 \leq j \leq N \tag{3.6}$$

Definidos os parâmetros N e $\lambda = \{A, B, \pi\}$, o HMM pode ser utilizado para gerar a sequência $O = \{O_1\, O_2\, O_3\, \ldots\, O_T\}$ (T é o número de observações na sequência) ou para

avaliar a verossimilhança de que uma dada sequência de observações tenha sido produzida pelo modelo.

Para que seja utilizado como classificador, calcula-se a probabilidade de uma dada sequência de observações $P(\mathbf{O}|\lambda_m)$. A classificação é determinada pela maior verossimilhança obtida entre as m classes. A probabilidade $P(\mathbf{O}|\lambda)$ é calculada por meio do algoritmo *Forward-Backward* [65]. A variável de avanço (*forward*) é definida por

$$\alpha_t(i) = P(O_1\, O_2\, O_3\, \ldots\, O_t, q_t = S_i \mid \lambda) \tag{3.7}$$

ou seja, a probabilidade da sequência de observação parcial até o instante t e que o estado seja S_i no tempo t, dado o modelo λ. A variável $\alpha_t(i)$ é inicializada por

$$\alpha_1(i) = \pi_i b_i(\mathbf{O}), \qquad 1 \leq i \leq N \tag{3.8}$$

e resolvida por indução, iterativamente, por

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(\mathbf{O}), \qquad 1 \leq t \leq T-1, \quad 1 \leq j \leq N, \tag{3.9}$$

fornecendo a quantidade desejada

$$P(\mathbf{O}|\lambda) = \sum_{i=1}^{N} \alpha_T(i). \tag{3.10}$$

A variável de avanço, por si só, é suficiente para se determinar a verossimilhança de uma sequência de observações $\mathbf{O}$, dado o modelo λ. A variável de retrocesso (*backward*) é complementar à de avanço e será utilizada no procedimento de estimação dos parâmetros do modelo. É definida como

$$\beta_t(i) = P(O_{t+1}\, O_{t+2}\, \ldots\, O_T, \mid q_t = S_i, \lambda) \tag{3.11}$$

Inicializando $\beta_t(i) = 1, \quad 1 \leq i \leq N$, e calculando por indução de forma decrescente em relação a t, conforme

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(\mathbf{O}) \beta_{t+1}(j), \quad t = T-1, T-2, \cdots, 1\,, \quad 1 \leq i \leq N, \tag{3.12}$$

Dado um conjunto de observações $\mathbf{O}$, os parâmetros de $\lambda = (A, B, \pi)$ podem ser estimados através de um procedimento iterativo de forma a maximizar $P(\mathbf{O}|\lambda)$. O algoritmo é conhecido como Baum-Welch, ou equivalentemente [65], o algoritmo EM – do inglês *Expectation-Maximization*.

Outras duas variáveis, além de $\alpha_t(i)$ e $\beta_t(i)$, serão necessárias para tal. A primeira é a probabilidade de se estar no estado S_i no instante t, dada a sequência de observações $\mathbf{O}$ e o modelo λ, ou seja,

$$\gamma_t(i) = P(q_t = S_i \mid \mathbf{O}, \lambda). \tag{3.13}$$

Em termos das variáveis de avanço e retrocesso, tem-se que

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)}. \tag{3.14}$$

Considerando, também, que as observações são contínuas, esta variável é definida como

$$\gamma_t(i,k) = \left[\frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)}\right]\left[\frac{c_{ik}\mathcal{N}(\mathbf{O}, \mu_{ik}, \mathbf{\Sigma}_{ik})}{\sum_{m=1}^{M} c_{ik}\mathcal{N}(\mathbf{O}, \mu_{ik}, \mathbf{\Sigma}_{ik})}\right] \tag{3.15}$$

onde $\gamma_t(i,k)$ é probabilidade de se estar no estado i no instante t levando em conta a componente da mistura k em $\mathbf{O}$.

A segunda variável denota a probabilidade de se estar no estado S_i em um dado instante t e transitar para o estado S_j no instante $t+1$, dados o modelo e a sequência de observações, ou seja,

$$\xi_t(i,j) = P(q_t = S_i,\, q_{t+1} = S_j \mid \mathbf{O}, \lambda). \tag{3.16}$$

Também pode ser calculada a partir das variáveis de avanço e retrocesso conforme

$$\xi_t(i,j) = \frac{\alpha_{ij} b_j(O_{t+1})\beta_{t+1}(j)}{\sum_{i=1^N} \sum_{j=1^N} \alpha_{ij} b_j(O_{t+1})\beta_{t+1}(j)} \tag{3.17}$$

A relação entre $\gamma_t(i)$ e $\xi_t(i,j)$ é dada por

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j) \tag{3.18}$$

Dessa forma, dado um modelo $\lambda = (A, B, \pi)$, os parâmetros podem ser atualizados conforme as seguintes equações:

$$\hat{\pi} = \gamma_1(i) \tag{3.19a}$$

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \tag{3.19b}$$

$$\hat{c}_{ik} = \frac{\sum_{t=1}^{T} \gamma_t(i,k)}{\sum_{t=1}^{T}\sum_{k=1}^{M} \gamma_t(i,k)} \tag{3.19c}$$

$$\hat{\mu}_{ik} = \frac{\sum_{t=1}^{T} \gamma_t(i,k) \cdot \mathbf{O}_t}{\sum_{t=1}^{T} \gamma_t(i,k)} \tag{3.19d}$$

$$\hat{\Sigma}_{ik} = \frac{\sum_{t=1}^{T} \gamma_t(i,k) \cdot (\mathbf{O}_t - \mu_{ik})(\mathbf{O}_t - \mu_{ik})^{\top}}{\sum_{t=1}^{T} \gamma_t(i,k)} \tag{3.19e}$$

3.3 Estimação de Dimensão Intrínseca

A partir do início da década de 60, a utilização de conjuntos de amostras (observações) caracterizadas por vetores para representar sinais, em lugar da representação por funções do tempo ou frequência, se tornou comum. Bennett [66] foi um dos primeiros a tentar definir a dimensionalidade de uma forma independente da escolha dos vetores de base que representam as amostras.

Segundo Bennett [66],

> A dimensão intrínseca de uma coleção de sinais é definida como o número de parâmetros livres requeridos por um gerador de sinais hipotético capaz de produzir uma boa aproximação de cada sinal no conjunto. Assim definida, a dimensão intrínseca se torna uma relação entre os vetores que representam os sinais. Esta relação não precisa ser linear e não depende da base na qual os vetores estão projetados. Ela representa um limite inferior no número de coeficientes necessários para descrever os sinais, não importando o quão sofisticado seja o esquema de representação, e assim provê um índice da redundância em uma dada representação.

Assim, um conjunto de dados $X \subset \mathcal{R}^D$ tem dimensão intrínseca (DI) $d \leq D$ se X puder ser descrito em termos de d graus de liberdade (parâmetros livres). O conjunto X é formado por vetores cujas componentes são funções de d variáveis aleatórias, $x_i = f_i(u_1, u_2, \ldots, u_d)$, $i = 1, 2, \ldots, D$, $u_i \in \mathcal{R}$. Pode-se interpretar geometricamente que o conjunto X está em uma hipersuperfície (*manifold* ou variedade) d-dimensional em $\mathcal{R}^D$ [67]

Nas tarefas de aprendizado de máquina e reconhecimento de padrões, a DI tem importância elevada, uma vez que um dos objetivos consiste na busca por fronteiras entre classes de dados. É provavelmente mais simples encontrá-las em espaços de dimensão

reduzida. Além disso, é bastante provável que um sinal não ocupe todo o hiperespaço $\mathcal{R}^D$. Por exemplo, os dígitos manuscritos da base de dados MNIST [52], cujas imagens têm dimensão 28×28 ou 1×784 (considerando a representação vetorial). Logo, estão num hiperespaço de representação $\mathcal{R}^{784}$ (tipicamente, o espaço de níveis de cinza é discreto, mas será assumido um espaço contínuo para simplificar as análises). A Figura 3.1 ilustra duas imagens contidas nesse hiperespaço.

(a) (b)

Figura 3.1 – Ilustração de imagens no hiperespaço 784−dimensional. Ilustração de um ponto (imagem) possível (a); Exemplo real do conjunto MNIST (b).

As duas imagens são pontos em $\mathcal{R}^{784}$, porém a Figura 3.1a aparenta ser apenas ruído, enquanto a Figura 3.1b é inteligível. Isto ilustra a intuição de que apenas uma parte do espaço é utilizada para a representação dos dígitos. A maioria dos *pixels*, principalmente os mais próximos às bordas, são sempre brancos em todos os exemplos do conjunto. Logo, os dados presentes na base MNIST, embora representados num espaço de alta dimensão, ocupam apenas uma pequena porção, uma variedade d-dimensional em $\mathcal{R}^{784}$.

Outro aspecto importante relacionado à DI é que o número de exemplos (amostras) necessários para representar satisfatoriamente as fontes de sinais cresce exponencialmente com o crescimento da dimensão dos mesmos – fenômeno chamado de *maldição da dimensionalidade* [67].

Por conta da importância da DI na análise de sinais e ferramentas nas áreas de reconhecimento de padrões e aprendizado de máquina, surgiram diversos métodos de estimação da dimensão intrínseca na literatura. O método utilizado neste trabalho para estimar a DI é inspirado em [68], e se preocupa particularmente com o caso do número de amostras ser pequeno.

Ele pode ser descrito da seguinte forma: primeiramente, assume-se que as observações são instâncias de uma variável aleatória multivariada, $\mathbf{X}$, com função densidade de probabilidade contínua, $f_{\mathbf{X}}(\mathbf{x})$. Assim, dados um conjunto de N amostras D-dimensionais cuja dimensão intrínseca é d ($d \leq D$), e uma região de observação com volume $V \propto r^D$ – onde r pode ser entendido como a largura dessa região em uma das D dimensões – então se r for pequeno o suficiente para que a densidade de probabilidade dentro de V seja constante, o número de pares de amostras dentro do volume deve aumentar proporcionalmente a r^d (ao invés de r^D).

Portanto, se $P(r)$ for a proporção de pares dentro do volume com relação à todos

os $N(N-1)/2$ pares de observações possíveis, e for aplicado logaritmo à aproximação $P(r) \approx K \cdot r^d$, a dimensão intrínseca pode ser estimada através da seguinte aproximação:

$$\log P(r) \approx d \cdot \log r + \log K$$

onde K é uma constante de proporcionalidade. Por isso, para estimar d, vários valores de r são testados, e o coeficiente angular da reta que melhor aproxima os pontos obtidos é utilizado como estimativa de d.

3.4 Codificação BRIEF

Descritores de pontos característicos são comuns em muitas aplicações de visão computacional como reconhecimento de objetos, reconstrução 3D, recuperação de imagem, entre outras aplicações, dos quais o SIFT [69] e o SURF [70] são dois exemplos bastante populares. Esses dois descritores são baseados em histograma de gradientes locais ao redor dos pontos de interesse da imagem, com a característica comum de serem invariantes à translação, escala e rotação da imagem. No entanto, o cálculo desses descritores tem uma carga computacional considerável.

O BRIEF [71] (*Binary Robust Independent Elementary Features*), é um descritor de pontos característicos de propósito geral baseado em sequências (*strings*) binárias computadas diretamente de fragmentos (*patches*) da imagem. Os *bits* individuais são obtidos através de testes de diferença de intensidade entre pares de *pixels*.

O descritor BRIEF é altamente discriminativo, mesmo usando um número relativamente pequeno de *bits*. A similaridade entre descritores BRIEF pode ser calculada com a distância de Hamming – em vez de distâncias comumente utilizadas, como a Euclidiana – que é eficientemente calculada através da operação lógica OU EXCLUSIVO (XOR) seguida pela contagem de *bits*.

Ao contrário do SIFT e SURF, o cálculo e a comparação entre vetores BRIEF é extremamente rápida e pouco custosa em termos de memória. Além disso, o BRIEF apresenta robustez a certos tipos de deformações fotométricas (iluminação) da imagem e tolera apenas pequenas rotações, sendo sensível à escala.

O procedimento de codificação BRIEF de uma imagem se dá da seguinte forma:

1. Seleciona-se aleatoriamente um conjunto de N pares de pontos $P_i = (\mathbf{x}_i, \mathbf{y}_i)$ na imagem;

2. Cria-se um vetor descritor $\mathbf{b} = \{b_1, b_2, \ldots, b_N\}$, inicialmente vazio.

3. Para cada par P_i, codifica-se segundo a regra:

$$b_i = \begin{cases} 1 & \text{se } I(\mathbf{x}_i) < I(\mathbf{y}_i) \\ 0 & \text{caso contrário} \end{cases}$$

onde $I(\mathbf{x})$ é a intensidade do *pixel* na posição $\mathbf{x}$.

4 Experimentos

O atual estado da arte no reconhecimento de fala por meio de leitura labial [14, 72, 73, 74] tem como essência as redes neurais, ferramenta que está novamente em voga devido à popularização do *deep learning*. De fato, a grande maioria dos trabalhos mais recentes se debruça sobre técnicas de aprendizado profundo, relegando técnicas consagradas, muitas das quais foram usadas, sobretudo a partir da segunda metade da década de 1990, desde o surgimento desse campo de pesquisa.

Por ser uma ferramenta consagrada no campo do reconhecimento de fala tradicional, o HMM logo foi transportado para o campo dos sinais visuais. O mapeamento entre fonemas e visemas permite a extensão das estruturas de reconhecimento de fala baseadas em HMM [9]. Apesar de ser uma ferramenta poderosíssima, os modelos ocultos de Markov dependem da extração de características relevantes. Por alguns anos, os trabalhos nessa área utilizaram modelos como ASM e AAM, entre outras técnicas, para extrair características visuais para os sistemas de leitura labial, que geralmente tinham o HMM como ferramenta para modelar as dependências temporais do sinal. Entretanto, essas técnicas de extração de características não eram tão robustas quanto às variações da pose do orador, luminosidade, mudança do orador etc.

Devido à grande disponibilidade de dados, evolução do poder computacional e a repopularização das redes neurais, o interesse dos trabalhos se voltou à utilização dessas redes não só para extração de características, como também para modelar as dependências temporais [3, 47, 48, 64]. Apesar dos resultados impressionantes, é muito difícil saber exatamente o que foi absorvido pelas camadas de uma rede neural convolucional (CNN), por exemplo, o que a rede aprendeu de informação quanto à fala ou quais características são de fato relevantes.

Recentemente, novos métodos de localização de pontos de referência faciais (*facial landmark detection*) têm surgido na comunidade científica [53, 55, 54, 56]. Não apenas mais robustos do que os AAMs, mas também menos sensíveis a amplas variações de iluminação, pose e, inclusive, tolerantes a oclusões parciais. O método descrito em [55] é usado como parte da arquitetura do atual estado da arte, para determinar as regiões de interesse nas imagens. Ademais, apesar da forte tendência quanto ao uso de redes neurais, o HMM – em menor proporção – ainda é uma ferramenta utilizada pelos pesquisadores no campo da leitura labial automática [75, 76].

Sendo assim, o interesse deste trabalho é utilizar pontos de referência faciais para extrair características relevantes para um sistema de leitura labial automática baseado em HMM. A extração de características é etapa fundamental e determinante em qualquer

sistema de classificação/reconhecimento de padrões. Aquelas devem ser adequadas ao classificador escolhido.

A reprodução de um trabalho da literatura foi o ponto inicial desta dissertação e motivou a discussão da adequabilidade dos vetores de características, sua dimensão intrínseca e a influência desses fatores no desempenho dos classificadores. Os experimentos descritos neste capítulo investigam estes fatos.

4.1 A base de dados Tulips1

A base de dados usada nos experimentos deste trabalho é a Tulips1 [1], publicamente disponível[1], e originalmente usada por Movellan [1] em experimentos sobre leitura labial.

A base Tulips1 é formada por amostras de áudio e vídeo separadas, coletadas de 12 voluntários (9 homens e 3 mulheres) do Departamento de Ciências Cognitivas na Universidade da Califórnia em San Diego – UCSD. A captura dos vídeos foi realizada em uma sala sem janelas no Centro de Pesquisas em Linguagem na UCSD. Os voluntários foram instruídos a falar, duas vezes, os quatro primeiros números em inglês para uma câmera de vídeo. Uma pequena tela foi posicionada em frente aos voluntários para que estes pudessem monitorar as imagens e se posicionar de maneira que seus lábios ficassem aproximadamente no centro da tela. As imagens foram digitalizadas em escala de cinza, a uma taxa de 30 *frames* por segundo (fps), 100×75 *pixels* e 8 *bits* por *pixel*. Os vídeos foram segmentados manualmente através da seleção de poucos *frames* relevantes antes e após o início e o fim das atividades registradas nas faixas de áudio [1].

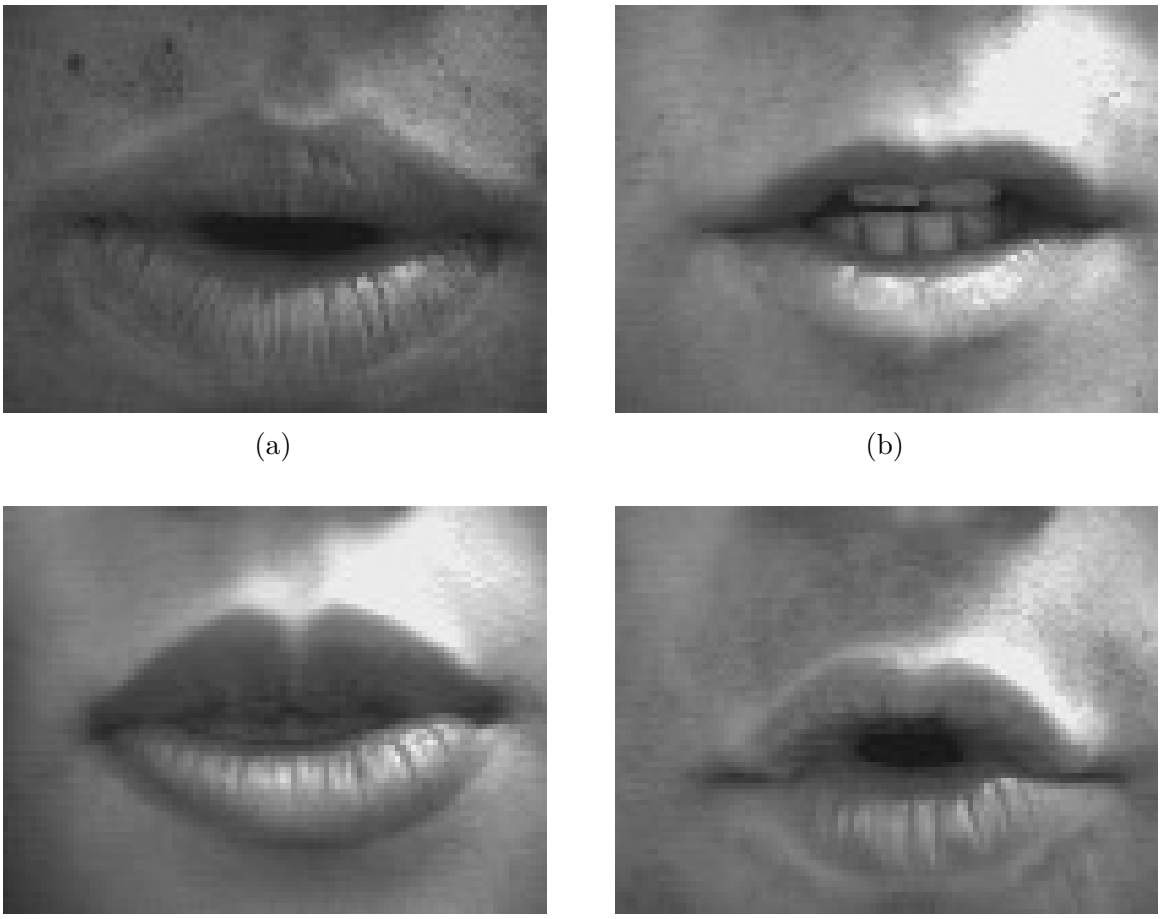
A base de dados contém um total de 934 imagens, os indivíduos possuem diferentes origens étnicas, alguns utilizam maquiagem ou possuem pêlos faciais e as condições de iluminação são diversas, apesar dos cuidados tomados durante a captura dos vídeos. As estatísticas das imagens do conjunto estão na Tabela 4.1.

Tabela 4.1 – Estatísticas da quantidade de *frames* por classe na base de dados Tulips1 [1].

Classe	Média	Desvio Padrão
'*one*'	8,9	2,1
'*two*'	9,6	2,1
'*three*'	9,7	2,3
'*four*'	10,6	2,2

Há, portanto, um total de 96 vídeos (12 indivíduos $\times$ 4 classes $\times$ 2 repetições). Na Figura 4.1, estão ilustradas 4 imagens aleatoriamente selecionadas nos vídeos, uma de cada classe existente na base de dados.

[1] <http://mplab.ucsd.edu/wordpress/?page_id=36>



(a) (b)

(c) (d)

Figura 4.1 – Ilustração de *frames* aleatórios das classes '*one*' (a), '*two*' (b), '*three*' (c) e '*four*' (d).

As imagens dessa base são controladas em escala e apresentam apenas a região da boca dos oradores, o que as tornam convenientes à realização de experimentos simples, em que a segmentação da região de boca não é necessária. Além disso, o criador dessa base a utilizou para realizar experimentos com 9 indivíduos, das quais 6 eram adultos com audição normal, sem treinamento em leitura labial e 3 adultos portadores de severas deficiências auditivas e treinados em leitura labial, estabelecendo assim uma referência de desempenho humano para experimentos de reconhecimento automático (89,93% e 95,49% de classificações corretas, respectivamente).

Como base de comparação, uma classificação aleatória deve produzir uma taxa de acertos médios de apenas 25%, ao passo que o desempenho humano obtido por 3 indivíduos treinados em reconhecimento labial foi de aproximadamente 95,5% de acurácia. Assim, espera-se que classificadores automáticos obtenham desempenhos dentro desse intervalo, o que de fato foi observado por Movellan e outros pesquisadores que também usaram essa mesma base [38, 77].

Sendo assim, as razões que levaram à escolha desta base para a realização dos

experimentos foram:

a) Ter sido utilizada em trabalhos na literatura que utilizaram HMM e algum tipo de modelo paramétrico, para fins de comparação com este trabalho;

b) O controle da posição do orador em relação à câmera;

c) Os vídeos já estão cortados e centrados na região de interesse (boca) dispensando assim a necessidade de se utilizar alguma técnica de reconhecimento da face e segmentação dos lábios.

d) O vocabulário pequeno é apropriado para a utilização de uma quantidade igualmente pequena de modelos do tipo HMM.

4.2 Reprodução dos Experimentos de J. R. Movellan

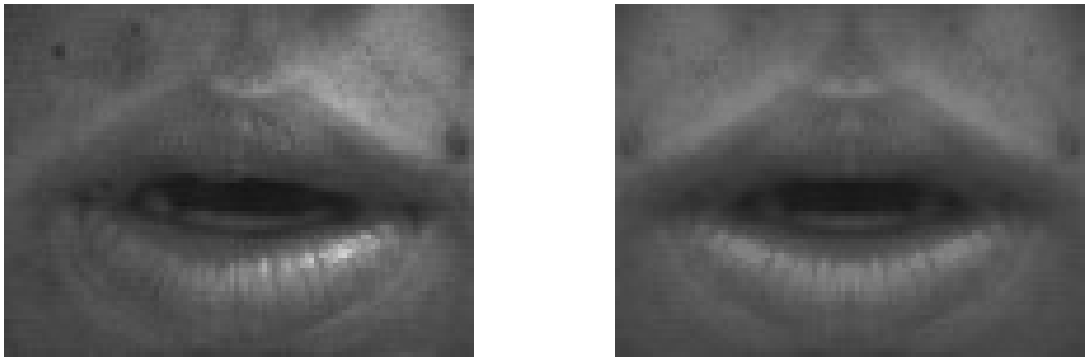
Como ponto de partida para este trabalho, decidiu-se reproduzir um trabalho encontrado na literatura para fins comparativos. Uma vez que o objetivo deste trabalho é entender como a redução do espaço de busca (através de uma boa escolha de características para o classificador) via de regra melhora o desempenho da tarefa de reconhecimento, nada mais apropriado do que realizar o reconhecimento das palavras pronunciadas utilizando técnicas menos rebuscadas (baseados nos níveis de cinza dos *pixels*) e a comparar com o resultado obtido quando se utiliza métodos mais elaborados (baseados em algum modelo).

Movellan [1] implementou um sistema reconhecedor baseado em HMM cujo treinamento foi realizado com os *bitmaps* em escala de cinza das imagens originais. O pré-processamento das imagens foi realizado através da sequência de etapas:

1) Simetrização das imagens: Utilizando a linha vertical no centro da imagem como eixo de simetria, cada *frame* foi simetrizado ao fazer a média *pixel* a *pixel* dos lados esquerdo e direito da imagem. O autor afirma que a simetrização introduz robustez e compressão, uma vez que o número de *pixels* é reduzido pela metade. Essas imagens são denominadas imagens-σ. A Figura 4.2 contém um exemplo da operação.

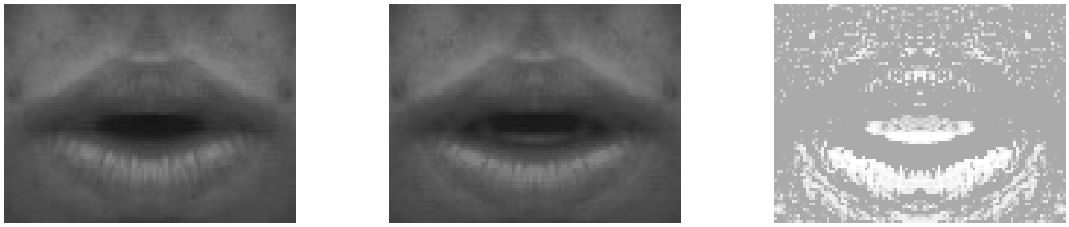
2) Diferenciação temporal: A cada instante de tempo são calculadas a diferença *pixel* a *pixel* entre o *frame* atual e o imediatamente anterior das imagens-σ. As potenciais vantagens dessa operação são a robustez às mudanças de iluminação e a ênfase aos aspectos dinâmicos do vídeo. As imagens resultantes são denominadas imagens-δ. A Figura 4.3 contém um exemplo da operação.

3) Filtragem gaussiana: As imagens-σ e imagens-δ foram filtradas por um filtro gaussiano, gerado por uma distribuição normal bidimensional $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ de média nula. O



(a) (b)

Figura 4.2 – Ilustração da operação de simetrização dos *frames*. Imagem original (a), imagem simetrizada (b).



(a) (b) (c)

Figura 4.3 – Ilustração da operação de diferenciação temporal dos *frames*. Imagem-σ no instante $t-1$ (a), imagem-σ no instante t (b) e imagem-δ no instante t (c). Apenas para fins de visualização, a imagem (c) teve seu histograma equalizado.

valor do desvio padrão σ_f do filtro que resultou em desempenho melhor, segundo Movellan, é igual a 4. A máscara do filtro (*kernel*) utilizada foi a seguinte:

$$\mathbf{F}_g = \begin{pmatrix} 0,1088 & 0,1123 & 0,1088 \\ 0,1123 & 0,1158 & 0,1123 \\ 0,1088 & 0,1123 & 0,1088 \end{pmatrix} \tag{4.1}$$

4) Subamostragem: Após a filtragem, as imagens-σ e imagens-δ resultantes foram subamostradas com um fator de 5, o que resultou em imagens com dimensões 20x15.

5) Limiarização: As imagens-σ e imagens-δ filtradas e subamostradas foram independentemente limiarizadas de acordo com a seguinte equação:

$$y = \frac{255}{1 + e^{\frac{-K\pi}{\sqrt{3\sigma}}(x-\mu)}}$$

onde μ e σ são a média e o desvio padrão das distribuições de níveis de cinza em cada sequência de imagens (vídeo). O valor da constante K (declividade da curva) que resultou no melhor desempenho final, segundo Movellan, é igual a 1,2.

6) Concatenação: As imagens-σ e imagens-δ resultantes foram divididas ao meio, em relação ao eixo vertical, e tiveram seus metades direitas e esquerdas, respectivamente, descartadas. As metades restantes foram concatenadas. As imagens resultantes possuem 300 *pixels* e são apresentadas ao HMM. A Figura 4.4 contém um exemplo do resultado final do pré-processamento das imagens, após a concatenação.

Figura 4.4 – Ilustração da imagem resultante após o pré-processamento de um *frame* aleatório.

Para modelar as sequências de imagens, foram implementados modelos de Markov do tipo *left-right* com 3 estados e uma mistura de 3 gaussianas por estado. Treinou-se um modelo HMM para cada classe na base de dados.

Assim como relatou o autor no artigo original, o treinamento do HMM utilizando o algoritmo EM sofreu diversos problemas numéricos. As soluções apontadas pelo autor para esses problemas foram as seguintes:

1. A restrição dos parâmetros de variância de todos os estados e misturas a um único valor, ou seja, uma única matriz de covariância e;
2. A inicialização dos centroides das misturas por segmentação linear e *K-means* [65, 78].

A segmentação linear divide as sequências de imagens de maneira uniforme entre os 3 estados, garantindo, assim, que os estados sejam treinados com aproximadamente a mesma quantidade de dados. Ademais, o algoritmo *K-means* estima os centroides das misturas, que outrora seriam inicializados aleatoriamente. Para cumprir a restrição de

utilizar os mesmos parâmetros de variância em todas as misturas e estados, uma matriz de covariância foi calculada a partir das imagens utilizadas no treinamento e se impôs uma restrição ao EM de realizar ajustes na mesma.

Entretanto, os problemas numéricos persistiram e, após análise, chegou-se às seguintes conclusões:

1. O número de amostras disponível é insuficiente para treinar modelos com tantos parâmetros (a matriz de covariância, por exemplo, possui dimensão 300×300;
2. Alguns dos parâmetros de variância calculados são nulos ou muito próximos de zero (devido ao fato de alguns *pixels*, principalmente nas bordas das imagens, permanecerem com os mesmos valores em diversas sequências inteiras de imagens).

Como solução, assumiu-se que os *pixels* são variáveis aleatórias descorrelacionadas, reduzindo, assim, a matriz de covariância a uma matriz diagonal. Além disso, as imagens foram normalizadas. Dessa forma, pôde-se utilizar a matriz identidade $\mathbf{I}$ em vez da matriz de covariância calculada a partir das imagens. Movellan não deixa claro se fez com que as covariâncias fossem nulas, apenas afirma que as variâncias dos *pixels* são iguais em todas as misturas e estados. Em relação à normalização dos dados e uso de matrizes diagonais, nada é mencionado em [1].

Como o conjunto de treinamento é pequeno, o teste se deu pelo procedimento *jackknife*, no qual os modelos foram treinados com 11 indivíduos, deixando um deles de fora para a validação. A taxa de generalização (acurácia[2]) encontrada foi de 64,58%, que é muito menor do que a relatada como o melhor resultado obtido pelo autor (89,58%). A Tabela 4.2 contém os resultados de classificação obtidos por Movellan.

Tabela 4.2 – Tabela de confusão dos resultados obtidos no trabalho original de Movellan [1].

-	**'one'**	**'two'**	**'three'**	**'four'**
'one'	100%	0%	0%	0%
'two'	4,17%	87,50%	4,17%	4,17%
'three'	12,5%	0%	83,33%	4,17%
'four'	8,33%	4,17%	0%	87,50%

O resultado do teste de reprodução dos resultados em [1] foi ligeiramente superior (66,67%) ao se relaxar a restrição das matrizes de covariância iguais para todos os estados e misturas. Foi permitido ao EM realizar o ajuste das matrizes de covariância, alterando apenas as variâncias de cada dimensão de forma a manter as matrizes diagonais, mas sem

[2] A acurácia dos resultados foi calculada a partir da soma dos elementos da diagonal principal da matriz de confusão dividida pela soma de todos os seu elementos.

permitir variâncias nulas. Cada padrão da classe i classificado como classe j foi dado como entrada na linha i e coluna j de uma matriz de confusão:

$$C = \begin{pmatrix} 20 & 0 & 3 & 1 \\ 4 & 13 & 6 & 1 \\ 3 & 3 & 18 & 0 \\ 0 & 1 & 10 & 13 \end{pmatrix} \tag{4.2}$$

A Tabela 4.3 contém os resultados em números percentuais para cada classe.

Tabela 4.3 – Tabela de confusão dos resultados obtidos na reprodução de [1].

-	**'one'**	**'two'**	**'three'**	**'four'**
'one'	83,33%	0%	12,5%	4,17%
'two'	16,67%	54,17%	25%	4,17%
'three'	12,5%	12,5%	75%	0%
'four'	0%	4,17%	41,67%	54,17%

O reconhecimento das classes foi, no mínimo, duas vezes superior à mera chance (25%), um indicativo de que o sistema aprendeu alguma informação visual acerca dos dígitos. De uma forma geral, os resultados da reprodução ficaram muito aquém daqueles obtidos no artigo original. Não só a acurácia do sistema reconhecedor (66,67% contra 89,58%), mas também o perfil dos acertos e erros de classificação em cada classe, conforme as Tabelas 4.2 e 4.3. Vale salientar que todos os algoritmos utilizados neste trabalho foram implementados, ou seja, não se usou bibliotecas e/ou pacotes computacionais publicamente disponíveis. Em [1] este aspecto não é mencionado.

Apesar dos resultados díspares entre a reprodução executada nesta dissertação e os relatados em [1], Movellan mostrou que é possível pré-processar e usar as imagens originais para a tarefa de reconhecimento de dígitos com HMM, mas o processo de ajuste não é robusto. As operações de pré-processamento originaram vetores de dimensões 1×300. São dimensões ainda muito altas, se comparadas à quantidade de exemplos, para que o sistema reconhecedor (HMM) seja capaz de aprender, satisfatoriamente, as características relevantes nos vídeos acerca da dinâmica labial. Se considerarmos que as imagens originais não têm dimensões grandes (100×75) e, ainda assim, os vetores de características gerados têm dimensões muito altas, não seria viável utilizar o mesmo sistema com uma base de dados mais recente, com imagens em alta definição, por exemplo, sem alguma operação de pré-processamento.

Além disso, é sabido que o algoritmo EM é problemático em relação à convergência, especialmente em espaços de alta dimensão. A redução da dimensão dos vetores de características é, não apenas bastante exequível, mas necessária para diminuir a possibilidade

do EM ficar preso em um mínimo local no hiperespaço dos parâmetros a serem ajustados e, ao contrário, convergir para o mínimo global.

4.3 Pontos de Referência Como Características para o HMM

4.3.1 Um Algoritmo de Rastreamento de Pontos Labiais Simples

Ainda que os seres humanos obtenham informações visuais das expressões faciais como um todo, a dinâmica labial, fundamentalmente, é a maior fonte de informações visuais na comunicação por fala, já que os lábios são uma estrutura proeminente do rosto. Dessa forma, há uma variedade de trabalhos que utilizam modelos paramétricos do contorno, forma e aparência dos lábios para extração de características [2, 18, 38, 75, 79]. Uma vez ajustado o modelo dos lábios na imagem, os parâmetros deste podem ser utilizados para o reconhecimento dos gestos labiais.

Para a etapa de extração de características, decidiu-se implementar, neste trabalho, uma estratégia simples: empregar apenas a posição de determinados pontos de referência marcados nos lábios para reconhecimento das palavras proferidas. Ainda que um algoritmo de localização de pontos faciais (*facial landmark detection*) mais rebuscado [53, 55] pudesse ser colocado em prática, a sua adequação aos vídeos desta base não seria simples. Primeiramente, ambos os algoritmos citados necessitam de um conjunto de treinamento muito maior do que o existente na Tulips1. Além disso, implementações de pacotes computacionais abertos como o OpenCV[3], por exemplo, partem do pressuposto de que a face precisa ser localizada na imagem para então localizar os pontos faciais de interesse (que abrangem toda a face e não somente a região da boca). Portanto, decidiu-se implementar um procedimento próprio para localização de pontos faciais.

A técnica é simples e requer a anotação dos pontos de interesse apenas no primeiro *frame* de cada vídeo na base Tulips1. Nos quadros seguintes, o algoritmo se encarrega de rastrear os pontos anteriormente marcados. Parte-se da premissa de que a iluminação permanece constante durante todo o vídeo e que o movimento labial é suave através dos *frames*.

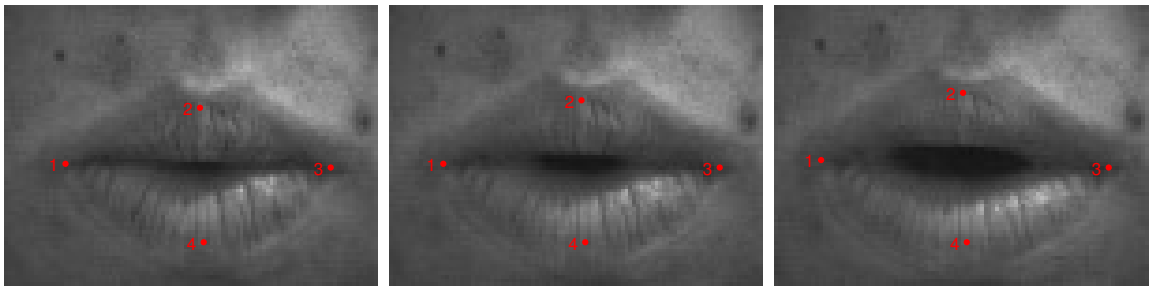
Dados os P_r, $r \in \{1, 2, \ldots, N_p\}$ pontos de referência em uma imagem (Figura 4.5 (a)), vetores descritores $\mathbf{V}_{r,t}$ da vizinhança de cada ponto são construídos. No *frame* subsequente, calculam-se os descritores $\mathbf{V}_{r,t+1}$ das coordenadas candidatas de cada ponto. A busca pelas novas coordenadas $(u_{r,t+1}, v_{r,t+1})$ é realizada dentro de uma janela de $L \times L$ *pixels* centrada em $(u_{r,t}, v_{r,t})$. As novas posições são determinadas por meio da comparação dos L^2 descritores candidatos $\mathbf{V}_{r,t+1}$ do *frame* atual e o descritor do *frame* imediatamente anterior $\mathbf{V}_{r,t}$. O algoritmo está descrito a seguir:

[3] <https://opencv.org/>

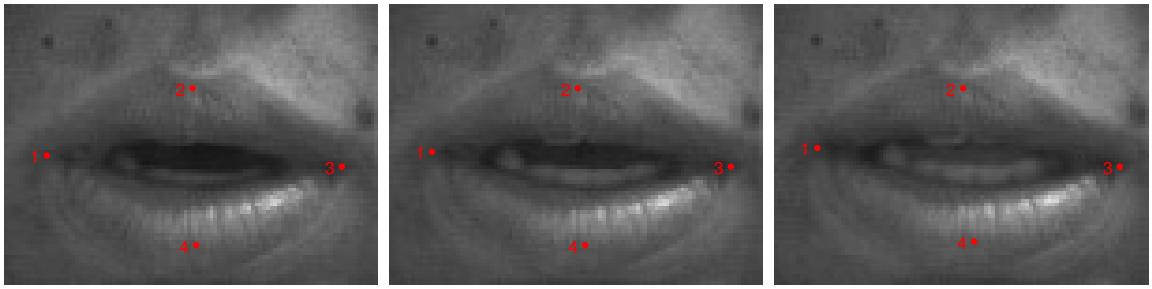
1. Escolhe-se, arbitrariamente, um número N_p de pontos de interesse a serem rastreados;
2. Marcam-se os P_r $r \in \{1, 2, \ldots, N_p\}$ pontos de interesse no primeiro *frame* do vídeo e suas coordenadas $(u_{r,0}, v_{r,0})$ são armazenadas;
3. Para todo P_r, um pequeno *patch* M_r de dimensões $L \times L$ centrado em $(u_{r,0}, v_{r,0})$ é definido;
4. Os *patches* M_r são vetorizados, têm sua média retirada e são normalizados, gerando os vetores descritores $\mathbf{V}_{r,0}$.
5. No *frame* seguinte ($t = 1$), para todo r, posiciona-se uma janela de $L \times L$ *pixels* centrada em $(u_{r,0}, v_{r,0})$, definindo, assim, uma margem de busca no entorno da última posição de cada P_r e se calculam os L^2 descritores candidatos $\mathbf{V}_{r,1}$ em cada posição da janela (i, j).
6. Os L^2 descritores candidatos $\mathbf{V}_{r,1}$ são comparados por meio de produto vetorial com os descritores do *frame* anterior $\mathbf{V}_{r,0}$. O máximo produto vetorial resultante está associado às coordenadas (i, j), e dessa maneira se atualizam as posições $(u_{r,1}, v_{r,1})$ dos pontos P_r. Armazenam-se a novas coordenadas dos pontos.
7. Os vetores descritores associados às novas coordenadas também são armazenados e utilizados na iteração seguinte. O procedimento segue desta forma até que o vídeo se encerre.
8. Para evitar que as buscas pelas coordenadas dos pontos ocorram fora das dimensões das imagens, foram impostas restrições quanto ao posicionamento das janelas próximo às bordas.

A Figura 4.5 ilustra o rastreamento de 4 pontos. Por inspeção visual, é possível notar que o algoritmo funciona satisfatoriamente, registrando as coordenadas dos pontos de interesse mesmo com a movimentação dos lábios. Entretanto, nem sempre o algoritmo consegue manter o controle sobre a localização dos pontos e perde a posição correta, passando a rastrear outro ponto cujo descritor é similar. Este fato ocorre basicamente em duas situações: quando o movimento dos lábios é muito rápido (na classe '*four*' isso é bastante evidente) e; quando os pontos de interesse se aproximam da borda da imagem e sofrem restrições na área de busca devido ao enquadramento inadequado, que pode ser causado pela movimentação inesperada da cabeça do orador. A Figura 4.6 contém um exemplo no qual o algoritmo perde o posicionamento correto do ponto de interesse. Além disso, a região mais interna dos lábios é naturalmente mais úmida, o que torna a pele mais reflexiva e influencia nas diferenças de iluminação entre as imagens.

É importante salientar que os pontos foram anotados manualmente no primeiro *frame* e o rastreamento foi realizado por uma técnica muito simples. Técnicas completamente automáticas e mais robustas quanto à variações de iluminação e variação de posição poderiam ter sido implementadas aqui. Porém, o objetivo desta etapa é apenas extrair características mais simples do que uma imagem pré-processada para alimentar o HMM.



(a) $t = 0$ (b) $t = 1$ (c) $t = 2$



(d) $t = 3$ (e) $t = 4$ (f) $t = 5$

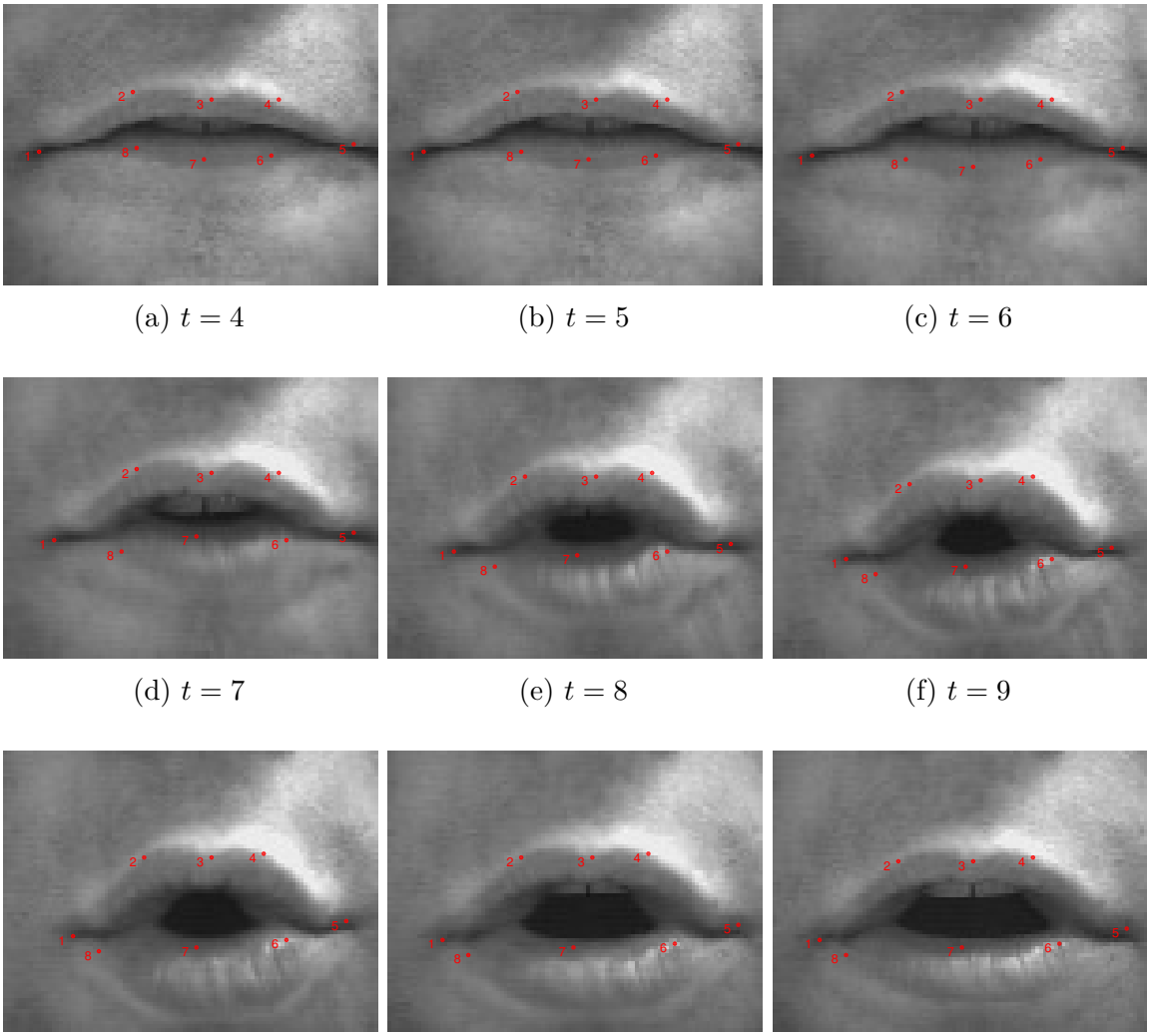
Figura 4.5 – Exemplo de vídeo da classe '*one*' onde é realizado o rastreamento dos pontos de interesse com $N_p = 4$.

4.3.2 Treinamento e Desempenho do HMM

É importante salientar que, diferentemente do ocorrido no experimento da Seção 4.2, as imagens utilizadas para o rastreamento dos pontos não foram pré-processadas. Decidiu-se pela utilização das imagens cruas, uma vez que o algoritmo de rastreamento exibiu bons resultados (na maioria dos vídeos) mesmo sem a suavização das imagens. O algoritmo gera um vetor associado a cada *frame* da seguinte forma:

$$\mathbf{S}_f = \begin{bmatrix} u_1 & v_1 & u_2 & v_2 & \cdots & u_{N_p} & v_{N_p} \end{bmatrix} \tag{4.3}$$

Em um primeiro momento, a acurácia na classificação dos dígitos utilizando os modelos HMM treinados com os vetores $\mathbf{S}_f$ brutos foi muito pior do que a obtida na reprodução do experimento de Movellan (Seção 4.2). Após uma avaliação criteriosa do problema, concluiu-se que as posições absolutas dos pontos de referência não é relevante na tarefa de classificação, mas apenas o do movimento, a dinâmica labial. Dessa forma, pôde-se remover as médias das componentes vertical (v) e horizontal (u) dos vetores $\mathbf{S}_f$,



(a) $t = 4$ (b) $t = 5$ (c) $t = 6$

(d) $t = 7$ (e) $t = 8$ (f) $t = 9$

(g) $t = 10$ (h) $t = 11$ (i) $t = 12$

Figura 4.6 – Exemplo de trecho de vídeo da classe '*four*' onde ocorre falha no rastreamento de alguns pontos de interesse com $N_p = 8$.

centralizando a boca na origem do sistema de coordenadas. Além disso, para retirar as influências das características inerentes a cada um dos oradores como o tamanho dos lábios e a amplitude dos movimentos labiais, os vetores foram normalizados, também em relação à cada componente do sistema de coordenadas. Essa configuração dos vetores de características levou aos melhores resultados em todas as configurações dos modelos de Markov.

Assim como na Seção 4.2, foram utilizados modelos HMM contínuos com distribuições gaussianas do tipo *left-right* (prática comum nas aplicações de reconhecimento de fala) e o treinamento foi realizado com o algoritmo EM. O procedimento utilizado também foi o de *jackknife*, no qual dados de 11 indivíduos foram usados para treinar e o restante para validação (*leave-one-out*), sendo este procedimento repetido 12 vezes, sempre com um indivíduo diferente para validação. A estimação inicial dos parâmetros do modelo HMM e das distribuições normais se deu por segmentação linear e *K-means*.

Novamente, o EM sofreu com problemas de convergência, uma vez que há muitos parâmetros livres para serem estimados e poucos dados para treinamento. Os problemas numéricos (devido ao surgimento de matrizes singulares) ocorrem mesmo com a imposição de restrições para evitar que elementos das matrizes de covariância tivessem elementos nulos, conforme sugere [65]. Para o número de pontos de referência rastreados $N_p = 4$, os vetores de características $\mathbf{S}_f$ têm dimensão 1×8, logo, as matrizes de covariância $\mathbf{R}$ têm dimensões 8×8 e, portanto, 64 parâmetros livres. Assumiu-se, então, que os pontos P_r são variáveis aleatórias independentes, fazendo com que as matrizes de covariância fossem diagonais.

O melhor resultado de classificação teve uma acurácia de 63,54% e foi obtido com modelos HMM com 4 estados e apenas uma gaussiana por estado. As matrizes de covariância foram inicializadas como matrizes identidade e não foi permitido o ajuste destes parâmetros pelo EM. A matriz de confusão do melhor resultado obtido (em valores absolutos) foi a seguinte:

$$C = \begin{pmatrix} 13 & 1 & 5 & 5 \\ 0 & 20 & 3 & 1 \\ 3 & 6 & 13 & 2 \\ 2 & 1 & 6 & 15 \end{pmatrix} \tag{4.4}$$

A matriz de confusão em termos percentuais pode ser vista na Tabela 4.4.

Tabela 4.4 – Tabela de confusão da classificação dos dígitos utilizando $N_p = 4$ pontos de referência como características.

-	**'one'**	**'two'**	**'three'**	**'four'**
'one'	54,17%	4,17%	20,83%	20,83%
'two'	0%	80,33%	12,5%	4,17%
'three'	12,5%	25%	54,17%	8,33%
'four'	8,33%	4,17%	25%	62,5%

De uma forma geral, o desempenho dos classificadores foi bastante próximo ao obtido no experimento da Seção 4.2 (66,67% e 63,54%). A maior diferença é que a estratégia de classificação com os pontos resultou em erros de classificação mais distribuídos entre classes, conforme uma inspeção das Tabelas 4.2 e 4.4 sugere. Ainda assim, todos os dígitos foram identificados corretamente em mais da metade dos testes. Em termos de tempo de treinamento, o uso dos vetores $\mathbf{S}_f$ é extremamente eficiente, da ordem de minutos, enquanto que o treinamento realizado na Seção 4.2 é da ordem de horas (para o mesmo computador e códigos implementados com mesmos cuidados relativos à eficiência computacional).

Supôs-se que aumentar as dimensões dos vetores de características para um valor intermediário entre as utilizadas nas duas estratégias até então adotadas elevaria o desempenho do classificador. Para testar a hipótese, foi gerado outro conjunto de dados de treinamento, agora com $N_p = 8$ pontos de referência. O objetivo era verificar a influência da quantidade de pontos rastreados no desempenho do classificador.

No caso de oito pontos de referência ($N_p = 8$), as matrizes de covariância foram inicializadas como matrizes identidade e também não foi permitido o ajuste destes parâmetros. O melhor resultado de classificação teve uma acurácia de 58,33% e foi obtido com modelos de Markov com 5 estados e apenas uma gaussiana por estado. A matriz de confusão do melhor resultado obtido (em valores absolutos) foi a seguinte:

$$C = \begin{pmatrix} 14 & 0 & 6 & 4 \\ 0 & 17 & 6 & 1 \\ 0 & 8 & 13 & 3 \\ 5 & 3 & 4 & 12 \end{pmatrix} \tag{4.5}$$

A matriz de confusão em termos percentuais pode ser vista na Tabela 4.5.

Tabela 4.5 – Tabela de confusão da classificação dos dígitos utilizando $N_p = 8$ pontos de referência como características.

-	**'one'**	**'two'**	**'three'**	**'four'**
'one'	58,33%	0%	25%	16,67%
'two'	0%	70,83%	25%	4,17%
'three'	0%	33,33%	54,17%	12,5%
'four'	20,83%	12,5%	16,67%	50%

O movimento dos pontos de referência P_r é correlacionado e o aumento indiscriminado na sua quantidade (além de um ponto crítico) não gera mais informações relevantes, pelo contrário, introduz redundância no sistema e torna o ajuste dos parâmetros do HMM mais difícil. Porém, não era de se esperar que o desempenho com o dobro de pontos ($N_p = 8$) fosse pior do que no experimento anterior. Ao se observar a Figura 4.5, é possível notar que, após a abertura total da boca a partir do quarto *frame*, as posições dos 4 pontos rastreados permanecem relativamente estáticas. Contudo, o contorno dos lábios continua a se deformar, sobretudo na parte interna dos mesmos. Era esperado que a inclusão de mais pontos fosse capaz de capturar essas informações mais finas quanto à dinâmica labial. Há ainda a possibilidade da insuficiência de dados para ajustar o modelo com mais parâmetros, por conta do pequeno tamanho da base de dados Tulips1.

4.4 Análise da Dimensão Intrínseca e Seu Efeito Na Classificação

O melhor resultado obtido na reprodução de Movellan foi de 66,67%, muito abaixo dos 89,58% obtidos no experimento original em [1]. Ademais, é interessante ressaltar que Movellan também realizou experimentos com HMM, porém sem a etapa de pré-processamento das imagens, o que resultou numa acurácia mais modesta de cerca de 55% de classificações corretas. Isto é um indicativo da provável inadequabilidade do HMM (*per si*) como modelo, considerando-se a desproporção entre números de parâmetros livres em cada HMM e o número de exemplos por classe.

Para testar essa hipótese e analisar o efeito da dimensão efetiva do espaço de representação dos dados sobre a acurácia dos classificadores, os experimentos foram realizados novamente, porém com esquemas de classificação bastante simplificados, a fim de que o efeito da dimensão não fosse compensado inadvertidamente pelos classificadores.

Em relação aos vetores de características, três alternativas foram empregadas na sua geração:

- A partir das imagens cruas da base;
- Empregando codificação BRIEF nas imagens e;
- A partir dos pontos de referência rastreados pelo algoritmo da Seção 4.3.

As seguintes diretrizes foram traçadas para os experimentos descritos nesta seção:

1) As imagens utilizadas não sofreram qualquer tipo de pré-processamento e/ou imposição de simetrias, ao contrário do que foi feito em [1].

2) Para representar a estrutura temporal dos gestos labiais, em lugar de um modelo estocástico como o HMM, uma codificação muito simples foi concebida, em que apenas 3 imagens concatenadas são usadas. Isto é, para cada vídeo representando um gesto, a primeira imagem do vídeo é concatenada à imagem no meio do vídeo, que finalmente é concatenada à última imagem do vídeo. A Figura 4.7 ilustra essa representação simplificada do gesto em 3 imagens concatenadas.

3) Na representação dos padrões, a mesma receita de simplicidade foi seguida, na qual cada gesto é codificado com a concatenação de três vetores, e cada vetor representa uma das três imagens codificadas. Portanto, cada gesto pode ser representado pela concatenação de três vetores com níveis de cinza dos *pixels*, representado por $\mathbf{g}_k$ para o k-ésimo vídeo, ou pela concatenação de três vetores com representações BRIEF de cada imagem, representado por $\mathbf{b}_k$, como ilustrado na Figura 4.7.

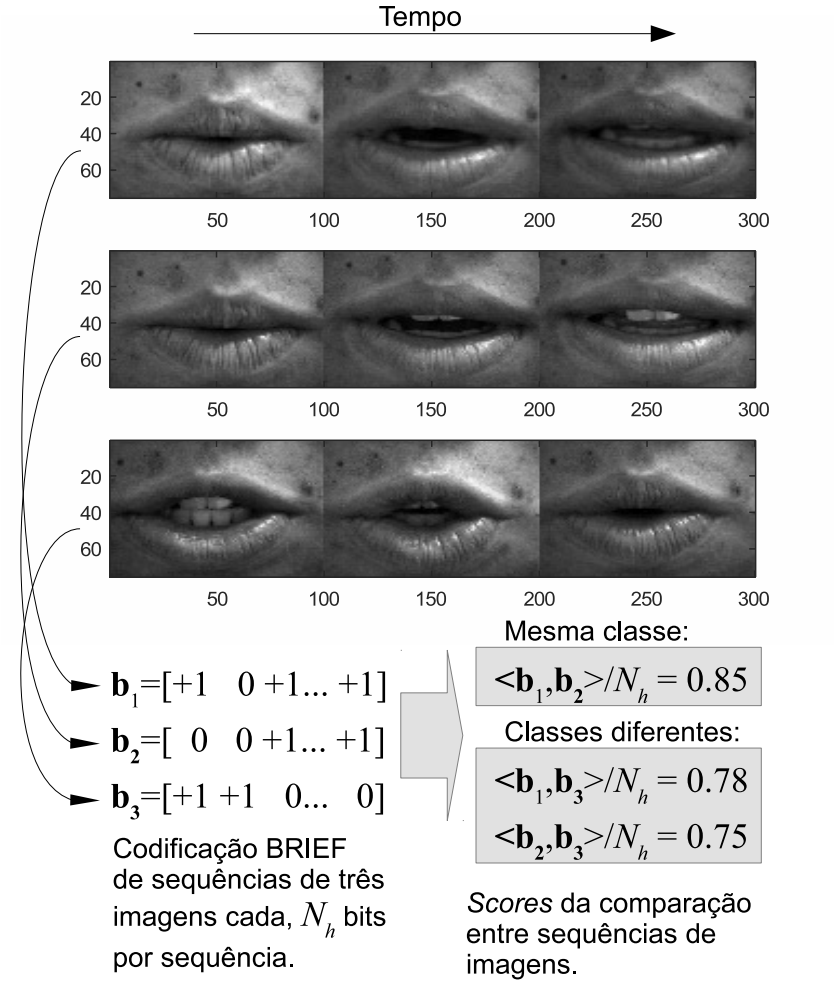


Figura 4.7 – Ilustração da codificação de gestos pela concatenação de (1) a imagem inicial do gesto, (2) a imagem no instante médio da duração do gesto e (3) a imagem final do gesto. Na parte superior da figura, três filmes da base são codificados com três imagens cada. As duas sequências superiores são rotuladas como gesto '*one*', enquanto a última é rotulada como gesto '*two*'. Cada imagem resultante é então codificada com BRIEF como um vetor $\mathbf{b}$ binário. No canto direito inferior da figura são ilustrados *scores* obtidos na comparação dos gestos através do produto escalar normalizado entre vetores binários, em que se pode perceber que os gestos de mesma classe pontuam mais alto, numa escala de 0 a 1, que os de classes diferentes.

4) No caso dos pontos de referência, em vez de se concatenar os 3 vetores de coordenadas $\mathbf{S}_f$ (Equação 4.3) correspondentes aos *frames* inicial, mediano e final do vídeo, cada gesto foi centralizado na origem e normalizado. Em seguida, foi dividido em 3 partes. As médias de cada terço foram concatenadas gerando, finalmente, os padrões $\mathbf{p}_k$ usados para classificação.

Dessa forma, haverá 96 padrões em escala de cinza $\mathbf{g}_k$, $k \in \{1, 2, \ldots, 96\}$, de dimensão 22500 ($3 \times 75 \times 100$), pois cada imagem possui 7500 *pixels* (75×100). Ou seja, 96 pontos definidos num espaço $\mathcal{R}^{22500}$. Usando o BRIEF, com o formato explicado na

seção 3.4, cada um desses 96 padrões pode ser convertido em um vetor binário, $\mathbf{b}_k$, com dimensão arbitrária. O número de comparações empregado foi igual a 3 vezes o número de *pixels* de cada imagem. Logo, cada vetor $\mathbf{b}$ é um vértice de um hipercubo no espaço $\{0,\ 1\}^{3\times 22500}$. Finalmente, haverá 96 padrões $\mathbf{p}_k$ de coordenadas espaciais dos pontos de referência, de dimensão 24 ($3 \times 4 \times 2$), média de 3 imagens com 4 pontos bidimensionais por imagem, ou dimensão 48 ($3 \times 8 \times 2$), a depender da quantidade de pontos de referência utilizados ($N_p = 4$ ou $N_p = 8$). Assim $\mathbf{g}_k$, $\mathbf{b}_k$ e $\mathbf{p}_k$ representam o mesmo gesto, mas em espaços de representação diferentes.

4.4.1 Dimensão Intrínseca de Imagens Concatenadas

A representação $\mathbf{g}$ é definida num espaço cuja dimensão é 22500. Se cada *pixel* da imagem fosse uma variável aleatória independente, o número de padrões necessários ao aprendizado (de máquina) para o reconhecimento dos dígitos seria muito maior que 96 (devido à maldição da dimensionalidade). Entretanto, os resultados mais modestos (55% de acurácia) em [1], obtidos sem o pré-processamento das imagens, estão significativamente acima da probabilidade de classificações aleatórias (25%), o que indica o aprendizado de alguma estrutura probabilística, mesmo com apenas 96 padrões em um espaço de alta dimensão nominal.

Logo, é evidente que os *pixels* não são independentes e, portanto, a dimensão intrínseca do espaço de características é muito menor do que a dimensão nominal. Para estimar a DI dos padrões $\mathbf{g}$, foi usada uma abordagem inspirada no estimador de DI de [68], vista na Seção 3.3.

A Figura 4.8 ilustra a análise feita na vizinhança do primeiro padrão (primeiro filme da base Tulips1). Como os eixos representam, respectivamente, distância e frequência relativa em escalas logarítmicas, espera-se que a DI se manifeste sob a forma de inclinação de uma reta. No entanto, o que se nota é o aparecimento de duas tendências lineares (duas retas, logo dois valores de DI possíveis).

A interpretação dada ao aparecimento das duas tendência é que a mais inclinada (maior DI), que aparece para vizinhanças menores do padrão, é devida a ruído (aditivo) das imagens, cuja intensidade é compatível com vizinhanças menores. Por essa razão, a primeira tendência foi descartada.

Repetindo essa análise para cada um dos 96 padrões, e calculando a média das 96 DIs encontradas, como sugerido em [68], foi estimado que as 96 concatenações de imagens formam uma estrutura cuja dimensão intrínseca é $DI \approx 14$. A DI também pode ser interpretada como uma medida do número de graus de liberdade da fonte geradora dos padrões. Portanto, os padrões $\mathbf{g}$ são tão difíceis de prever quanto uma imagem com 14 *pixels* independentes, o que dá uma ideia do nível de dificuldade imposto a um classificador

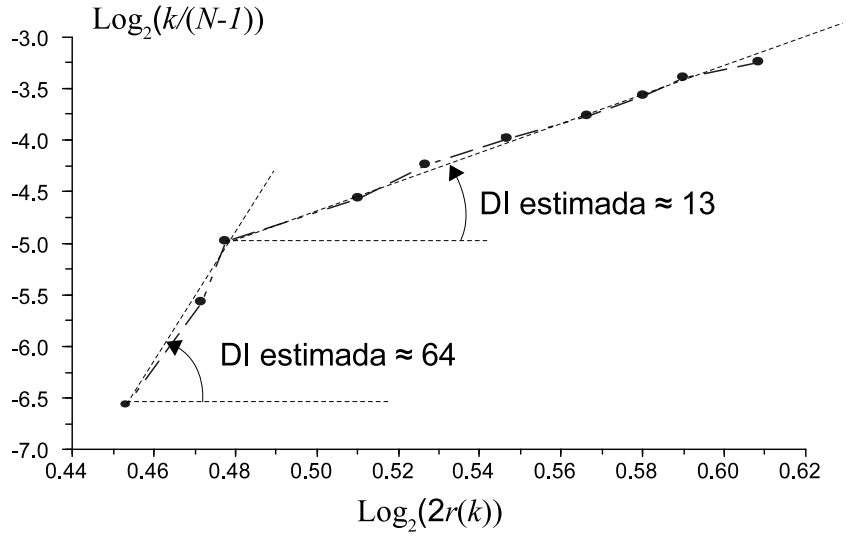


Figura 4.8 – Ilustração da análise da DI para padrões gerados a partir de três imagens codificadas em níveis de cinza. A ilustração é feita apenas na vizinhança do padrão 1 – que pode ser visto na Figura 4.7 –, com base numa vizinhança de 10 padrões mais próximos. As distâncias ordenadas de forma crescente são representadas como $r(k)$, onde k indica a proximidade ao padrão 1 (i.e. $r(k)$ indica a distância, em norma infinita, do k-ésimo padrão mais próximo).

de sequências de imagens, codificadas diretamente como vetores de níveis de cinza $\mathbf{g}$.

4.4.2 Dimensão Intrínseca de Padrões Codificados com BRIEF

Até onde o autor deste texto conhece a literatura relacionada ao tema, o conceito de DI aplicado a padrões representados com código BRIEF ainda não foi estudado. Portanto, o método desenvolvido nesta subseção é a maior contribuição desta dissertação.

Embora a DI seja geralmente associada a uma variável contínua, num espaço de representação $\mathcal{R}^D$, a definição de Bennett (Seção 3.3) pode ser estendida às variáveis discretas. Se um padrão for visto como uma instância de variável aleatória discreta, a definição de Bennett remete ao conceito de entropia da fonte, na medida em que o número mínimo de variáveis independentes e uniformes capazes de gerar as observações é a própria entropia da fonte das amostras [80].

Para ilustrar essa conexão entre entropia e dimensão intrínseca, suponhamos que se tem pequenas imagens 2×2 cujos níveis de cinza dos *pixels* sejam modelados por quatro variáveis aleatórias dependentes, $\mathrm{X}_{1,1}$, $\mathrm{X}_{1,2}$, $\mathrm{X}_{2,1}$ e $\mathrm{X}_{2,2}$, geradas por duas variáveis aleatórias independentes, uniformes (no intervalo $[0, 1]$) e latentes Z_1 e Z_2, conforme ilustrado na Figura 4.9. Além disso, 25 instâncias dessa fonte aleatória também foram apresentadas.

Após exame das instâncias, é possível inferir alguma dependência entre os quatro *pixels*, o que visualmente revela que há menos de quatro graus de liberdade. A análise das instâncias de $\mathbf{X}$ diretamente no seu espaço de representação original contínuo em 4D pelo

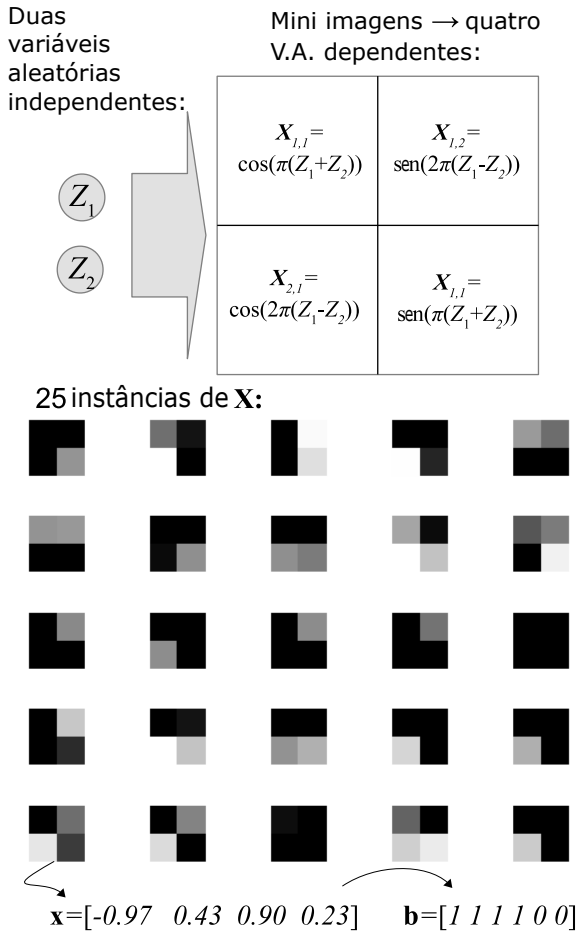


Figura 4.9 – Ilustração do modelo gerador de pequenas imagens aleatórias 2×2. 25 (5×5) instâncias das matriz aleatória correspondente, $\mathbf{X}$, foram apresentadas. Além disso, o exemplo no canto inferior esquerdo foi apresentado numericamente, com sua respectiva codificação BRIEF (exaustiva).

estimador de DI discutido na Seção 3.3 infere que haja apenas 2 graus de liberdade no modelo gerador, ou seja, apenas duas fontes independentes.

Ao se codificar as imagens exaustivamente com BRIEF, onde todos os seis possíveis pares de *pixels* são comparados, em vez dos pares aleatórios usados na codificação BRIEF comum, cada instância é associada a uma sequência de seis *bits*, $\mathbf{b}$, como ilustrado na Figura 4.9. Logo, cada vetor $\mathbf{b}$ pode ser considerado uma instância de uma variável aleatória discreta $\mathbf{B}$.

Para esta ilustração do conceito, realizou-se um experimento com 1000 instâncias independentes de $\mathbf{B}$. Então, ao redor de cada instância, todas as instâncias vizinhas não mais distantes do que 1 (em distância de Hamming) foram agrupadas como um subconjunto das instâncias. A entropia de Shannon associada a esse subconjunto foi estimada de maneira direta, utilizando a frequência relativa das instâncias vizinhas, e resultou numa entropia local estimada de cerca de 1,82 *bits* (com desvio padrão empírico de 0,01, em 100 replicações independentes deste experimento).

Portanto, de acordo com a definição de DI de Bennett mencionada acima, 1,82 é a estimativa das instâncias de **B**. Assim como era esperada uma estimativa de DI próxima de 2 para **X**, também se espera que a entropia ou o número de graus de liberdade local de **B** seja próximo a 2. Ambas as estimativas (de entropia local e dimensão intrínseca) remetem ao número de graus de liberdade da fonte geradora. Os resultados mostram que a entropia local pode ser utilizada como estimativa de DI no espaço de representação do BRIEF. A codificação BRIEF não é uma função invertível de **X** em **B**, o que implica em perda de informação, porém o seu uso adequado deve preservar informações relevantes.

Para representações BRIEF de imagens reais, o uso direto da frequência relativa para estimar a entropia local conforme demonstrado no exemplo anterior não é razoável devido à dispersão dos dados, ou seja, pontos na vizinhança de um determinado código BRIEF têm baixa probabilidade de aparecerem mais de uma vez. Para superar esta limitação, um modelo gerador simples foi utilizado para as reais instâncias de **B**. Este modelo é fundamentado nas seguintes premissas:

P1: A distância de Hamming entre padrões BRIEF indica diferença entre imagens comparadas.

P2: Os N_b *bits* de um padrão BRIEF são determinados por N_z variáveis latentes independentes e binárias.

P3: Cada variável latente binária determina aproximadamente N_b/N_z *bits* no código BRIEF.

Graças a P1, é possível se definir uma vizinhança de um dado padrão BRIEF, $\mathbf{b}_k$, como a coleção de V padrões mais próximos, em distância de Hamming. Se as premissas P2 e P3 forem válidas, as distâncias desses V vizinhos mais próximos não serão discrepantes, e seu valor médio será N_b/N_z, donde se pode estimar N_z, que é o número de graus de liberdade na vizinhança de $\mathbf{b}_k$.

A Figura 4.7 ilustra o uso de P1 na geração de *scores* (que é uma premissa do próprio BRIEF), enquanto que a Figura 4.10 ilustra a estimação de N_z para a primeira sequência de 3 imagens da base (a sequência de imagens do topo da Figura 4.7), para uma vizinhança de $V = 10$ padrões mais próximos, que é a mesma vizinhança usada na estimação de DI (ilustrado na Figura 4.8). Repetindo esse procedimento para os 96 padrões da base, para codificações BRIEF, foi obtida uma estimativa praticamente constante (robusta) de menos de 5 graus de liberdade. Isso se manteve mesmo quando a codificação BRIEF foi alterada, apenas neste experimento, de 67500 ($3 \times 75 \times 100 \times 3$) a 225000 ($3 \times 75 \times 100 \times 10$) *bits* por padrão.

Usando um raciocínio comparativo sobre os significados da DI e dos graus de liberdade, os padrões **b** são tão difíceis de prever quanto uma imagem binária com menos

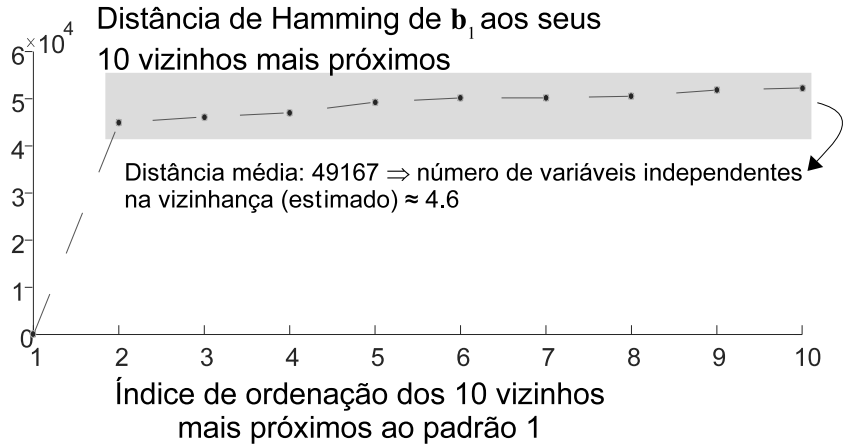


Figura 4.10 – Ilustração da análise do número de graus de liberdade de padrões codificadas em BRIEF. A ilustração é feita apenas na vizinhança do padrão 1, com base numa vizinhança de 10 padrões mais próximos.

de 5 *pixels* independentes, o que, ao contrário do que acontece com os padrões $\mathbf{g}$, é claramente compatível com um aprendizado/treinamento de máquina baseado em apenas 96 exemplos.

4.4.3 Dimensão Intrínseca dos Padrões de Coordenadas dos Pontos de Referência

Das três representações definidas no início da Seção 4.4, $\mathbf{p}_k$ é a que tem menor dimensão nominal, sendo definida nos espaços $\mathcal{R}^{24}$ ou $\mathcal{R}^{48}$, para um número de pontos de referência rastreados (N_p), que pode ser, respectivamente, igual a 4 ou 8.

Utilizando o método de estimação descrito na Seção 3.3, estimou-se a DI a partir dos 10 vizinhos mais próximos. Para o caso de $N_p = 4$, os padrões formam um conjunto cuja dimensão intrínseca estimada é $DI \approx 6$. Já para o caso de $N_p = 8$, $DI \approx 9$.

4.4.4 Resultados de Classificação

Como alternativa ao HMM, que apresentaram os problemas de excesso de parâmetros livres para a base em uso aqui, escolheu-se utilizar classificadores lineares [67], conforme a equação:

$$y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0 = 0 \tag{4.6}$$

onde $y(\mathbf{x})$ é o hiperplano de decisão entre as classes no espaço D-dimensional, $\mathbf{w} = [w_1, w_2, \ldots, w_D]^T$ é o vetor de pesos e w_0 é uma constante. Os coeficientes w_j foram ajustados pelo método dos mínimos quadrados e pseudoinversão. Uma implementação para cada tipo de representação, cada espécie de vetores de características ($\mathbf{g}_k$, $\mathbf{b}_k$ e $\mathbf{p}_k$) foi realizada, com as seguintes especificações:

- Projetando os 96 padrões **g** em representações *one-hot* da classe associada ao padrão. Por exemplo, a classe '*one*' é representada como [1 0 0 0], enquanto a classe '*four*', como [0 0 0 1];
- Projetando os 96 padrões **b** também em representações *one-hot* da classe associada ao padrão;
- Projetando os 96 padrões **p** (para $N_p = 4$ e $N_p = 8$) em representações *one-hot* da classe associada ao padrão.

As projeções obtidas foram classificadas de acordo com a posição do valor máximo no vetor de saída. O procedimento de validação foi o mesmo *jackknife* usado em [1].

No caso da codificação BRIEF, que é baseada em gerador de números pseudoaleatórios, cada experimento foi repetido 20 vezes, fornecendo as taxas de confusões médias apresentadas na Tabela 4.6.

Tabela 4.6 – Tabela de confusões para classificação linear com BRIEF. As linhas representam as classes verdadeira e as colunas, as classes decididas pelo classificador. O elemento na linha i e coluna j representa o número médio de padrões da classe i atribuído à classe j.

-	**'one'**	**'two'**	**'three'**	**'four'**
'one'	20,3	1,65	1	1,05
'two'	0	22	0	2
'three'	1	4	15,8	3,2
'four'	0,3	1,2	0,65	21,85

Todas as acurácias dos experimentos realizados nesta Seção, além dos resultados obtidos com HMM nas Seções 4.2 e 4.3 foram apresentadas na Tabela 4.7. Além dos resultados de classificação, foram incluídas as dimensões nominal e intrínseca calculadas para fins de comparação. Como o BRIEF é uma codificação que possui uma componente aleatória na escolha do código, o seu resultado é apresentado em termos de média e desvio padrão para 20 repetições independentes do experimento com códigos diferentes.

Novamente, o resultado da classificação com HMM para imagens brutas foi de aproximadamente 55% em [1]. A representação baseada na concatenação de 3 imagens (padrões **g**) pode ser considerada equivalente ao uso de imagens não processadas. Ao se comparar aquele resultado com o obtido na linha 2 da Tabela 4.7, fica evidente a melhora de desempenho pelo emprego de vetores de características com DI reduzida e a importância do processo de extração de características.

Tabela 4.7 – Resultados comparativos do melhor classificador HMM com os classificadores lineares dos padrões **g**, **b** e **p**.

Método	**Acurácia**	**Dim. Nominal**	**DI**
HMM com vídeo (de 6 a 16 imagens pré-processadas p/ vídeo)	66,67%	300	—
Classif. linear com 3 imagens concatenadas em níveis de cinza (**g**)	77%	22500	≈ 14
Classif. linear com 3 imagens concatenadas em cod. BRIEF (**b**)	83%± 1,0%	67500	≈ 5
Classif. linear com rastreamento de $N_p = 4$ pontos de referência (**p**)	81,25%	24	≈ 6
Classif. linear com rastreamento de $N_p = 8$ pontos de referência (**p**)	71,88%	48	≈ 9

4.5 Discussão dos Resultados

Como a base de dados usada neste trabalho apresenta uma relação de desproporção acentuada entre dimensão nominal de representação de padrões (D) e número de padrões disponíveis para treinamento (N), é esperado um impacto destacado da DI – ou do número de graus de liberdade da fonte geradora dos sinais (padrões) – sobre a acurácia de classificadores.

Theodoridis e Koutroumbas [67] afirmam que, geralmente, a probabilidade de erros de classificação para um determinado classificador é mínima quando a proporção $\alpha = N/D$ está no intervalo $[2, 10]$. A regra, embora abranja uma vasta porção de classificadores ditos tradicionais, não é válida para qualquer classificador, por exemplo, o SVM (*Support Vector Machine*). Nos casos em que N é finito, não é o número de parâmetros que, de fato, controla o desempenho da generalização, mas outra quantidade, a qual, para alguns tipos de classificadores, está relacionada ao número de parâmetros a serem estimados e também à dimensão intrínseca do espaço de representação. Isso fica bastante evidente ao analisar a capacidade de generalização do classificador linear de padrões **b**, que, embora tenha o maior número de parâmetros livres para serem ajustados, trabalha, também, com a menor DI dos conjuntos de representações usados.

De fato, a DI dos padrões **g** indica que há aproximadamente 14 fontes independentes a serem modeladas, e não se deve esperar que isso seja feito de forma satisfatória com apenas 96 exemplos. Essa análise dá também uma ideia aproximada do que acontece com os vídeos, que são composições de mais de 3 imagens (entre 6 e 16 imagens por vídeo[4]). Logo, não devem ser gerados com menos graus de liberdade que aqueles associados aos padrões **g**. Isso condiz com as dificuldades de ajuste dos modelos HMM e com a necessidade de imposição de inúmeras etapas de pré-processamento às imagens, além das restrições ao

[4] A dimensão nominal dos padrões brutos varia de 45000 a 120000 (7500 *pixels* × nº de *frames* no vídeo).

treinamento dos modelos estocásticos (reduzindo o número de graus de liberdade em dois momentos distintos) para melhorar o desempenho do classificador de dígitos.

Nesse sentido, a construção dos padrões **g** também representa uma redução de graus de liberdades em relação aos vídeos, o que impacta positivamente no desempenho de um classificador tão simples quanto o linear. Mas a DI dos padrões **g** ainda é muito elevada para a quantidade de padrões disponíveis para treinamento, e o classificador linear também termina tendo um papel relevante de restrição de liberdades no momento da classificação. Para testar esse efeito, um classificador do tipo *k Nearest Neighbour* (k-NN) [67] foi usado no lugar do classificador linear dos padrões **g**, sob mesmo protocolo de validação, e o melhor resultado, com 3-NN, foi uma acurácia de 51%, que é comparável ao desempenho do HMM para imagens sem pré-processamento.

A codificação BRIEF foi responsável por uma redução intensa do número de graus de liberdade, como esperado, pois os níveis de cinza dos *pixels* são descartados, substituídos pelo resultado da mera comparação entre intensidades entre pares de *pixels*. Essa redução para aproximadamente 5 graus de liberdade é corroborada pelo incremento de acurácia do classificador linear para taxas compatíveis com os melhores resultados com HMM. Usando-se um classificador 1-NN (melhor valor de k neste caso) no lugar do classificador linear dos padrões **b**, sob mesmo protocolo de validação, o resultado se degrada para uma acurácia de 57%, o que confirma a importância também dem as restrições impostas à fronteira de classificação pelo classificador linear, mas confirma a vantagem que o BRIEF provê ao descartar mais da metade dos graus de liberdade dos padrões **g**. Isso é compatível com o descarte de subespaços ocupados por sinais ruidosos ou irrelevantes.

Finalmente, a caracterização dos gestos labiais pelo movimento de pontos de referência (padrões **p**) também reduziu consideravelmente a dimensão intrínseca dos dados. Embora as DIs obtidas não tenham sido melhores do que aquela associada à codificação BRIEF, as dimensões nominais desses vetores foram muito menores (24 e 48) do que em todos os outros experimentos, resultando em melhores generalizações. No caso de $N_p = 4$, o desempenho é comparável aos obtidos com o uso do BRIEF. Aqui também foi realizada a mesma análise de substituição do classificador linear por um k-NN, sob o mesmo protocolo de validação. Para $N_p = 4$, um 9-NN resultou numa acurácia de 77,08% e, no caso de $N_p = 8$, um 1-NN atingiu a taxa de generalização de 66,67%. Assim como ocorrido com as representações **g** e **b**, houve degradação (não tão acentuada neste caso) do desempenho com o k-NN para as duas quantidades de pontos rastreadas.

A acurácia obtida pelos pontos de referência como vetores de características no HMM foi de 63,54% e 58,33%, para quatro e oito pontos rastreados, respectivamente. Ambos são inferiores ao alcançado com uso das imagens pré-processadas. Ademais, o classificador linear se mostrou muito sensível às pequenas alterações na estrutura de **p**, apresentando resultados da ordem de 50% de acerto (esses resultados foram obtidos

durante experimentações com estruturas de **p** diversas das reportadas na Seção 4.4). É de se esperar que o HMM, mais complexo do que o classificador linear, tenha tido tantas dificuldades para ajustar um modelo satisfatório nessa representação que descarta uma enorme quantidade de informação.

5 Conclusões

Nesta dissertação foi investigada a relação da dimensão intrínseca dos vetores de características e o desempenho de classificação em um sistema de leitura labial. As dificuldades na reprodução do trabalho de Movellan [1] motivaram a elaboração de melhores vetores de características e o estudo da influência da dimensão intrínseca dos mesmos na acurácia de classificadores.

Na etapa inicial, de reprodução do trabalho de Movellan, o treinamento dos HMMs sofreu com problemas de convergência. Movellan justificou o desempenho ruim da classificação com o argumento de que as densidades de probabilidade das imagens rapidamente tendiam a zero devido à alta dimensão dos vetores. De fato, este problema também ocorreu na reprodução conduzida. Impuseram-se restrições nos parâmetros de variância e inicialização do treinamento, conforme o trabalho original, mas os problemas persistiram. Uma análise mais profunda mostrou que a própria natureza dos dados de treinamento (imagens preprocessadas) era a causa. Algumas das dimensões dos vetores de características, especialmente aquelas referentes aos *pixels* mais próximos às bordas, tinham variâncias nulas ou muito próximas de zero, resultando em singularidades. A solução encontrada foi a normalização dos dados e utilização de matrizes de covariância identidades. Supõe-se que, embora essas restrições impostas no treinamento tenham permitido a convergência e sanado os problemas numéricos, a estrutura probabilística dos dados foi alterada dificultando o reconhecimento do conjunto de testes. Movellan não menciona problemas similares.

A hipótese de que ocorreu sobreajuste (*overfitting*) no treinamento foi descartada. O procedimento de validação consistia em testar, por exemplo, as amostras do indivíduo 1 no modelo da classe '*one*' treinado sem aquele mesmo indivíduo. Dessa forma, para cada classe, geraram-se 12 modelos ao final. Testando-se as amostras do indivíduo 1 nos 11 modelos que as incluíam no treinamento, mas excluíam a de outra pessoa, não se obteve resultados significativamente melhores, indicando que não houve sobreajuste.

Esperava-se que a utilização de pontos de referência reduzisse não só a dimensão nominal dos vetores de características (em relação aos vídeos), como também a DI – já que $0 \leq d \leq D$ –, melhorando o desempenho do HMM. De fato, a dimensão nominal foi drasticamente reduzida, teoricamente possibilitando um ajuste satisfatório dos modelos, dado o número de amostras do conjunto de treinamento. Entretanto, os resultados não foram significativamente diferentes (considerando apenas 4 pontos de interesse). Embora as DIs dos vídeos e das trajetórias de pontos não tenham sido estimadas, o desempenho insatisfatório se deu, provavelmente, por conta das múltiplas restrições que tiveram que

ser mantidas por conta dos mesmos problemas de convergência identificados ou, ainda, por ter utilizado menos dimensões do que a DI, descartando inadvertidamente informações relevantes.

Para mostrar que a DI realmente influencia nas acurácias, foram propostos esquemas de classificação alternativos baseados classificadores lineares e vetores de características concatenados. A concatenação de vetores é uma forma (muito simples) de estimar a trajetória dos sinais no decorrer do tempo. Os vetores **g** constituem uma maneira análoga ao uso das imagens brutas e sem processamento como características para o HMM. A análise da DI dos vetores **g** – que certamente é menor ou igual à DI de sequências inteiras de imagens brutas – mostrou que o número de exemplos na base de dados é incompatível com bons desempenhos do HMM.

Em relação aos vetores **b**, mostrou-se que a DI não está atrelada à dimensão nominal dos vetores de base, mas intimamente relacionada aos conceitos de entropia e graus de liberdade. De todos os vetores de características avaliados, apresentaram a menor DI e, conforme esperado, os melhores resultados de classificação.

Já os resultados obtidos com os vetores **p** (considerando apenas 4 pontos rastreados) confirmaram que a redução da DI se reflete no desempenho de classificação. Basta comparar as taxas de generalização de **g** e **p**. No caso de 8 pontos rastreados, os resultados sugerem a ocorrência do clássico fenômeno de pico (*peaking phenomenon*) [67], ou seja, o aumento do número de características na tentativa de se diminuir a taxa de erros, mas que resulta exatamente no oposto por conta do número limitado de amostras no conjunto de treinamento.

Deve-se salientar que os classificadores lineares são muito simples e via de regra não são utilizados para modelar fenômenos com dependências temporais. Muito provavelmente obteriam desempenhos inferiores em uma base de dados com vocabulário maior, o que não pôde ser testado neste trabalho. Além disso, também restringem graus de liberdade, o que é refletido nas acurácias obtidas, a julgar pelos resultados obtidos com classificadores k-NN.

Finalmente, os vetores de características usados neste trabalho são frutos da investigação de características adequadas ao problema da leitura labial automática. Conforme discutido na introdução deste texto, ainda não há um consenso na comunidade científica sobre essa questão. Para trabalhos futuros, propõe-se a inclusão de outros tipos de informação nos vetores de características, como derivadas e informações sobre luminosidade e textura no entorno dos lábios, além da utilização de um conjunto de dados mais amplo.

Referências

1 MOVELLAN, J. R. Visual speech recognition with stochastic networks. In: *Advances in neural information processing systems*. [S.l.: s.n.], 1995. p. 851–858. Citado 12 vezes nas páginas 5, 6, 9, 22, 32, 34, 37, 38, 45, 47, 52 e 56.

2 MATTHEWS, I. et al. Extraction of visual features for lipreading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 24, n. 2, p. 198–213, Feb 2002. ISSN 0162-8828. Citado 6 vezes nas páginas 13, 15, 16, 19, 22 e 39.

3 WAND, M.; KOUTNÍK, J.; SCHMIDHUBER, J. Lipreading with long short-term memory. *CoRR*, abs/1601.08188, 2016. Disponível em: <http://arxiv.org/abs/1601.08188>. Citado 4 vezes nas páginas 13, 20, 22 e 31.

4 POTAMIANOS, G. et al. The handbook of multimodal-multisensor interfaces. In: OVIATT, S. et al. (Ed.). New York, NY, USA: Association for Computing Machinery and Morgan & Claypool, 2017. cap. Audio and Visual Modality Combination in Speech Processing Applications, p. 489–543. ISBN 978-1-97000-167-9. Disponível em: <https://doi.org/10.1145/3015783.3015797>. Citado 4 vezes nas páginas 13, 14, 21 e 23.

5 MCGURK, H.; MACDONALD, J. Hearing lips and seeing voices. Nature Publishing Group, 1976. Citado na página 13.

6 MACDONALD, J.; MCGURK, H. Visual influences on speech perception processes. *Perception & Psychophysics*, v. 24, n. 3, p. 253–257, 1978. ISSN 1532-5962. Disponível em: <http://dx.doi.org/10.3758/BF03206096>. Citado na página 13.

7 EPHRAT, A.; PELEG, S. Vid2speech: Speech reconstruction from silent video. *arXiv preprint arXiv:1701.00495*, 2017. Citado na página 13.

8 FISHER, C. G. Confusions among visually perceived consonants. *Journal of Speech, Language, and Hearing Research*, v. 11, n. 4, p. 796–804, 1968. Disponível em: <+http://dx.doi.org/10.1044/jshr.1104.796>. Citado na página 14.

9 HARTE, N.; GILLEN, E. Tcd-timit: An audio-visual corpus of continuous speech. *IEEE Transactions on Multimedia*, v. 17, n. 5, p. 603–615, May 2015. ISSN 1520-9210. Citado 2 vezes nas páginas 14 e 31.

10 DAVIS, K. H.; BIDDULPH, R.; BALASHEK, S. Automatic recognition of spoken digits. *The Journal of the Acoustical Society of America*, v. 24, n. 6, p. 637–642, 1952. Disponível em: <http://dx.doi.org/10.1121/1.1906946>. Citado na página 14.

11 OLSON, H. F.; BELAR, H. Phonetic typewriter. *The Journal of the Acoustical Society of America*, v. 28, n. 6, p. 1072–1081, 1956. Disponível em: <http://dx.doi.org/10.1121/1.1908561>. Citado na página 14.

12 FORGIE, J. W.; FORGIE, C. D. Results obtained from a vowel recognition computer program. *The Journal of the Acoustical Society of America*, v. 31, n. 11, p. 1480–1489, 1959. Disponível em: <http://dx.doi.org/10.1121/1.1907653>. Citado na página 14.

13 JUANG, B.-H.; RABINER, L. R. Automatic speech recognition – a brief history of the technology development. *Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara*, v. 1, p. 67, 2005. Citado na página 14.

14 CHUNG, J. S. et al. Lip reading sentences in the wild. *CoRR*, abs/1611.05358, 2016. Disponível em: <http://arxiv.org/abs/1611.05358>. Citado 3 vezes nas páginas 15, 20 e 31.

15 LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. *nature*, Nature Publishing Group, v. 521, n. 7553, p. 436, 2015. Citado 2 vezes nas páginas 15 e 21.

16 POTAMIANOS, G. et al. Audio-visual automatic speech recognition: An overview. *Issues in visual and audio-visual speech processing*, MIT Press Cambridge, MA, v. 22, p. 23, 2004. Citado 2 vezes nas páginas 15 e 17.

17 ONG, E. J.; BOWDEN, R. Robust facial feature tracking using shape-constrained multiresolution-selected linear predictors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 33, n. 9, p. 1844–1859, Sept 2011. ISSN 0162-8828. Citado 2 vezes nas páginas 15 e 19.

18 BOWDEN, R. et al. Recent developments in automated lip-reading. In: INTERNATIONAL SOCIETY FOR OPTICS AND PHOTONICS. *SPIE Security+ Defence*. [S.l.], 2013. p. 89010J–89010J. Citado 3 vezes nas páginas 15, 19 e 39.

19 ZHOU, Z. et al. A review of recent advances in visual speech decoding. *Image and Vision Computing*, v. 32, n. 9, p. 590 – 605, 2014. ISSN 0262-8856. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0262885614001036>. Citado na página 16.

20 MONTALVÃO, J.; ARAUJO, M. R. R. Is masking a relevant aspect lacking in mfcc? a speaker verification perspective. *Pattern Recogn. Lett.*, Elsevier Science Inc., New York, NY, USA, v. 33, n. 16, p. 2156–2165, dez. 2012. ISSN 0167-8655. Disponível em: <http://dx.doi.org/10.1016/j.patrec.2012.07.023>. Citado na página 16.

21 PETAJAN, E.; GRAF, H. P. Automatic lipreading research: Historic overview and current work. In: _____. *Multimedia Communications and Video Coding*. Boston, MA: Springer US, 1996. p. 265–275. ISBN 978-1-4613-0403-6. Disponível em: <http://dx.doi.org/10.1007/978-1-4613-0403-6_33>. Citado na página 17.

22 DENBY, B. et al. Silent speech interfaces. *Speech Communication*, v. 52, n. 4, p. 270 – 287, 2010. ISSN 0167-6393. Silent Speech Interfaces. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0167639309001307>. Citado na página 17.

23 PETAJAN, E. et al. An improved automatic lipreading system to enhance speech recognition. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 1988. (CHI '88), p. 19–25. ISBN 0-201-14237-6. Disponível em: <http://doi.acm.org/10.1145/57167.57170>. Citado 2 vezes nas páginas 17 e 18.

24 HASEGAWA, T.; OHTANI, K. Oral image to voice converter-image input microphone. In: *[Proceedings] Singapore ICCS/ISITA '92*. [S.l.: s.n.], 1992. p. 617–620 vol.2. Citado na página 17.

25 YUHAS, B. P. et al. Neural network models of sensory integration for improved vowel recognition. *Proceedings of the IEEE*, v. 78, n. 10, p. 1658–1668, Oct 1990. ISSN 0018-9219. Citado 2 vezes nas páginas 17 e 18.

26 WU, J.-T. et al. Neural network vowel-recognition jointly using voice features and mouth shape image. *Pattern Recognition*, v. 24, n. 10, p. 921 – 927, 1991. ISSN 0031-3203. Disponível em: <http://www.sciencedirect.com/science/article/pii/003132039190089N>. Citado na página 18.

27 STORK, D. G.; WOLFF, G.; LEVINE, E. Neural network lipreading system for improved speech recognition. In: *[Proceedings 1992] IJCNN International Joint Conference on Neural Networks*. [S.l.: s.n.], 1992. v. 2, p. 289–295 vol.2. Citado na página 18.

28 BREGLER, C. et al. Improving connected letter recognition by lipreading. In: *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*. [S.l.: s.n.], 1993. v. 1, p. 557–560 vol.1. ISSN 1520-6149. Citado na página 18.

29 GOLDSCHEN, A. J.; GARCIA, O. N.; PETAJAN, E. Continuous optical automatic speech recognition by lipreading. In: *Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers*. [S.l.: s.n.], 1994. v. 1, p. 572–577 vol.1. ISSN 1058-6393. Citado na página 18.

30 BREGLER, C.; KONIG, Y. Eigenlips for robust speech recognition. In: *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*. [S.l.: s.n.], 1994. ii, p. II/669–II/672 vol.2. ISSN 1520-6149. Citado na página 18.

31 KASS, M.; WITKIN, A.; TERZOPOULOS, D. Snakes: Active contour models. *International journal of computer vision*, Springer, v. 1, n. 4, p. 321–331, 1988. Citado 3 vezes nas páginas 18, 19 e 22.

32 COOTES, T. F.; TAYLOR, C. J. et al. *Statistical models of appearance for computer vision*. 2004. Citado na página 19.

33 CHIOU, G. I.; HWANG, J.-N. Lipreading from color motion video. In: *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*. [S.l.: s.n.], 1996. v. 4, p. 2156–2159 vol. 4. ISSN 1520-6149. Citado na página 19.

34 CHIOU, G. I.; HWANG, J.-N. Lipreading from color video. *IEEE Transactions on Image Processing*, v. 6, n. 8, p. 1192–1195, Aug 1997. ISSN 1057-7149. Citado na página 19.

35 COOTES, T. F.; TAYLOR, C. J. Active shape models — 'smart snakes'. In: _____. *BMVC92: Proceedings of the British Machine Vision Conference, organised by the British Machine Vision Association 22–24 September 1992 Leeds*. London: Springer London, 1992. p. 266–275. ISBN 978-1-4471-3201-1. Disponível em: <http://dx.doi.org/10.1007/978-1-4471-3201-1_28>. Citado na página 19.

36 LUETTIN, J.; THACKER, N. A.; BEET, S. W. Active shape models for visual speech feature extraction. In: _____. *Speechreading by Humans and Machines: Models, Systems, and Applications*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1996. p. 383–390. ISBN 978-3-662-13015-5. Disponível em: <http://dx.doi.org/10.1007/978-3-662-13015-5_28>. Citado na página 19.

37 LUETTIN, J.; THACKER, N. A.; BEET, S. W. Visual speech recognition using active shape models and hidden markov models. In: *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*. [S.l.: s.n.], 1996. v. 2, p. 817–820 vol. 2. ISSN 1520-6149. Citado na página 19.

38 LUETTIN, J.; THACKER, N. A. Speechreading using probabilistic models. *Computer Vision and Image Understanding*, v. 65, n. 2, p. 163 – 178, 1997. ISSN 1077-3142. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1077314296905706>. Citado 3 vezes nas páginas 19, 33 e 39.

39 COOTES, T. F.; EDWARDS, G. J.; TAYLOR, C. J. Active appearance models. In: _____. *Computer Vision — ECCV'98: 5th European Conference on Computer Vision Freiburg, Germany, June 2–6, 1998 Proceedings, Volume II*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998. p. 484–498. ISBN 978-3-540-69235-5. Disponível em: <http://dx.doi.org/10.1007/BFb0054760>. Citado na página 19.

40 COOTES, T. F.; EDWARDS, G. J.; TAYLOR, C. J. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 23, n. 6, p. 681–685, Jun 2001. ISSN 0162-8828. Citado 2 vezes nas páginas 19 e 22.

41 COOTES, T. F. et al. Comparing active shape models with active appearance models. In: *Bmvc*. [S.l.: s.n.], 1999. v. 99, n. 1, p. 173–182. Citado na página 19.

42 ONG, E.-J.; BOWDEN, R. Robust lip-tracking using rigid flocks of selected linear predictors. In: *2008 8th IEEE International Conference on Automatic Face&Gesture Recognition*. [S.l.: s.n.]. Citado na página 19.

43 ONG, E. J. et al. Robust facial feature tracking using selected multi-resolution linear predictors. In: *2009 IEEE 12th International Conference on Computer Vision*. [S.l.: s.n.], 2009. p. 1483–1490. ISSN 1550-5499. Citado na página 19.

44 LAN, Y. et al. Comparing visual features for lipreading. In: *International Conference on Auditory-Visual Speech Processing 2009*. [S.l.: s.n.], 2009. p. 102–106. Citado na página 19.

45 LAN, Y.; HARVEY, R.; THEOBALD, B. J. Insights into machine lip reading. In: *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.: s.n.], 2012. p. 4825–4828. ISSN 1520-6149. Citado na página 19.

46 BOWDEN, R. et al. Is automated conversion of video to text a reality? In: . [s.n.], 2012. v. 8546, p. 85460U–85460U–9. Disponível em: <http://dx.doi.org/10.1117/12.979437>. Citado na página 19.

47 GRAVES, A. et al. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In: *Proceedings of the 23rd International Conference on Machine Learning*. New York, NY, USA: ACM, 2006. (ICML '06), p. 369–376. ISBN 1-59593-383-2. Disponível em: <http://doi.acm.org/10.1145/1143844.1143891>. Citado 2 vezes nas páginas 19 e 31.

48 NODA, K. et al. Lipreading using convolutional neural network. In: . [S.l.: s.n.], 2014. Citado 3 vezes nas páginas 19, 22 e 31.

49 COOKE, M. et al. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, ASA, v. 120, n. 5, p. 2421–2424, November 2006. Disponível em: <http://link.aip.org/link/?JAS/120/2421/1>. Citado na página 20.

50 VIOLA, P.; JONES, M. Rapid object detection using a boosted cascade of simple features. In: *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001.* [S.l.: s.n.], 2001. v. 1, p. I–I. ISSN 1063-6919. Citado na página 21.

51 DALAL, N.; TRIGGS, B. Histograms of oriented gradients for human detection. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05).* [S.l.: s.n.], 2005. v. 1, p. 886–893 vol. 1. ISSN 1063-6919. Citado na página 21.

52 LECUN, Y. et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, v. 86, n. 11, p. 2278–2324, Nov 1998. ISSN 0018-9219. Citado 2 vezes nas páginas 21 e 28.

53 CAO, X. et al. Face alignment by explicit shape regression. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition(CVPR).* [s.n.], 2012. v. 00, p. 2887–2894. ISSN 1063-6919. Disponível em: <doi.ieeecomputersociety.org/10.1109/CVPR.2012.6248015>. Citado 3 vezes nas páginas 21, 31 e 39.

54 CAO, C.; HOU, Q.; ZHOU, K. Displaced dynamic expression regression for real-time facial tracking and animation. *ACM Trans. Graph.*, ACM, New York, NY, USA, v. 33, n. 4, p. 43:1–43:10, jul. 2014. ISSN 0730-0301. Disponível em: <http://doi.acm.org/10.1145/2601097.2601204>. Citado 2 vezes nas páginas 21 e 31.

55 KAZEMI, V.; SULLIVAN, J. One millisecond face alignment with an ensemble of regression trees. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition.* [S.l.: s.n.], 2014. p. 1867–1874. ISSN 1063-6919. Citado 3 vezes nas páginas 21, 31 e 39.

56 FENG, Y. et al. Joint 3d face reconstruction and dense alignment with position map regression network. *CoRR*, abs/1803.07835, 2018. Disponível em: <http://arxiv.org/abs/1803.07835>. Citado 2 vezes nas páginas 21 e 31.

57 LU, Y.; YAN, J.; GU, K. Review on automatic lip reading techniques. *International Journal of Pattern Recognition and Artificial Intelligence*, v. 32, n. 07, p. 1856007, 2018. Citado 3 vezes nas páginas 21, 22 e 23.

58 GRITZMAN, A. D.; RUBIN, D. M.; PANTANOWITZ, A. Comparison of colour transforms used in lip segmentation algorithms. *Signal, Image and Video Processing*, v. 9, n. 4, p. 947–957, May 2015. ISSN 1863-1711. Disponível em: <https://doi.org/10.1007/s11760-014-0615-x>. Citado na página 21.

59 WOLFF, G. J. et al. Lipreading by neural networks: Visual preprocessing, learning and sensory integration. In: *Proceedings of the 6th International Conference on Neural Information Processing Systems.* San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993. (NIPS'93), p. 1027–1034. Disponível em: <http://dl.acm.org/citation.cfm?id=2987189.2987318>. Citado na página 22.

60 HECKMANN, M. et al. Dct-based video features for audio-visual speech recognition. In: *7th International Conference on Spoken Language Processing, ICSLP2002 - INTERSPEECH 2002, Denver, Colorado, USA, September 16-20, 2002*. [s.n.], 2002. Disponível em: <http://www.isca-speech.org/archive/icslp_2002/i02_1925.html>. Citado na página 22.

61 POTAMIANOS, G.; GRAF, H. P.; COSATTO, E. An image transform approach for hmm based automatic lipreading. In: *Proceedings 1998 International Conference on Image Processing. ICIP98 (Cat. No.98CB36269)*. [S.l.: s.n.], 1998. p. 173–177 vol.3. Citado na página 22.

62 MASE, K.; PENTLAND, A. Automatic lipreading by optical-flow analysis. *Systems and Computers in Japan*, Wiley Subscription Services, Inc., A Wiley Company, v. 22, n. 6, p. 67–76, 1991. ISSN 1520-684X. Disponível em: <http://dx.doi.org/10.1002/scj.4690220607>. Citado na página 22.

63 KOLLER, O.; NEY, H.; BOWDEN, R. Deep learning of mouth shapes for sign language. In: *The IEEE International Conference on Computer Vision (ICCV) Workshops*. [S.l.: s.n.], 2015. Citado na página 22.

64 PAREKH, D. et al. Lip reading using convolutional auto encoders as feature extractor. *CoRR*, abs/1805.12371, 2018. Disponível em: <http://arxiv.org/abs/1805.12371>. Citado 2 vezes nas páginas 22 e 31.

65 RABINER, L. R. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, v. 77, n. 2, p. 257–286, Feb 1989. Citado 4 vezes nas páginas 23, 25, 36 e 43.

66 BENNETT, R. *Representation and analysis of signals?Part XXI: The intrinsic dimensionality of signal collections*. [S.l.], 1965. Citado na página 27.

67 THEODORIDIS, S.; KOUTROUMBAS, K. *Pattern Recognition, Fourth Edition*. 4th. ed. [S.l.]: Academic Press, 2008. ISBN 1597492728, 9781597492720. Citado 6 vezes nas páginas 27, 28, 51, 53, 54 e 57.

68 FARAHMAND, A. m.; SZEPESVáRI, C.; AUDIBERT, J.-Y. Manifold-adaptive dimension estimation. In: *Proceedings of the 24th International Conference on Machine Learning*. New York, NY, USA: ACM, 2007. (ICML '07), p. 265–272. ISBN 978-1-59593-793-3. Citado 2 vezes nas páginas 28 e 47.

69 LOWE, D. G. Object recognition from local scale-invariant features. In: IEEE. *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*. [S.l.], 1999. v. 2, p. 1150–1157. Citado na página 29.

70 BAY, H.; TUYTELAARS, T.; GOOL, L. V. Surf: Speeded up robust features. In: SPRINGER. *European conference on computer vision*. [S.l.], 2006. p. 404–417. Citado na página 29.

71 CALONDER, M. et al. Brief: Binary robust independent elementary features. In: SPRINGER. *European conference on computer vision*. [S.l.], 2010. p. 778–792. Citado na página 29.

72 CHUNG, J. S.; ZISSERMAN, A. Lip reading in the wild. In: *Asian Conference on Computer Vision*. [S.l.: s.n.], 2016. Citado na página 31.

73 CHUNG, J. S.; ZISSERMAN, A. Out of time: automated lip sync in the wild. In: *Workshop on Multi-view Lip-reading, ACCV*. [S.l.: s.n.], 2016. Citado na página 31.

74 CHUNG, J. S.; ZISSERMAN, A. Learning to lip read words by watching videos. *Computer Vision and Image Understanding*, 2018. ISSN 1077-3142. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1077314218300134>. Citado na página 31.

75 KOLLER, O.; NEY, H.; BOWDEN, R. Read my lips: Continuous signer independent weakly supervised viseme recognition. In: SPRINGER. *European Conference on Computer Vision*. [S.l.], 2014. p. 281–296. Citado 2 vezes nas páginas 31 e 39.

76 REKIK, A.; BEN-HAMADOU, A.; MAHDI, W. Human machine interaction via visual speech spotting. In: *Proceedings of the 16th International Conference on Advanced Concepts for Intelligent Vision Systems - Volume 9386*. Berlin, Heidelberg: Springer-Verlag, 2015. (ACIVS 2015), p. 566–574. ISBN 978-3-319-25902-4. Disponível em: <https://doi.org/10.1007/978-3-319-25903-1_49>. Citado na página 31.

77 MATTHEWS, I. et al. A comparison of active shape model and scale decomposition based features for visual speech recognition. In: _____. *Computer Vision — ECCV'98: 5th European Conference on Computer Vision Freiburg, Germany, June 2–6, 1998 Proceedings, Volume II*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998. p. 514–528. ISBN 978-3-540-69235-5. Disponível em: <http://dx.doi.org/10.1007/BFb0054762>. Citado na página 33.

78 FINK, G. A. *Markov Models for Pattern Recognition: From Theory to Applications*. 2nd. ed. [S.l.]: Springer Publishing Company, Incorporated, 2014. ISBN 1447163079, 9781447163077. Citado na página 36.

79 ROTHKRANTZ, L. Lip-reading by surveillance cameras. In: *2017 Smart City Symposium Prague (SCSP)*. [S.l.: s.n.], 2017. p. 1–6. Citado na página 39.

80 SHANNON, C. E. A mathematical theory of communication. *The Bell System Technical Journal*, v. 27, n. 3, p. 379–423, July 1948. ISSN 0005-8580. Citado na página 48.